SASCHA HAUKE

ON THE STATISTICS OF TRUSTWORTHINESS PREDICTION

ON THE STATISTICS OF TRUSTWORTHINESS PREDICTION

SASCHA HAUKE

Vom Fachbereich Informatik
der Technischen Universität Darmstadt genehmigte

Dissertation

zur Erlangung des akademischen Grades

Doctor rerum naturalium (Dr. rer. nat.)

Eingereicht von:
Dipl.-Inform. Sascha Hauke
geboren in Recklinghausen

Erstreferent: Prof. Dr. Max Mühlhäuser
Korreferent: Prof. Christian D. Jensen, Ph.D.
Tag der Einreichung: 14. Januar 2015
Tag der Prüfung: 16. März 2015

Fachgebiet Telekooperation
Fachbereich Informatik
Technische Universität Darmstadt
Hochschulkennziffer D-17

Darmstadt 2015

Sharon: *How do you know? I mean, how do you really know that you can trust me?*
Adama: *I don't. That's what trust is.*
— *Battlestar Galactica*, "Precipice" (2006)

Take everything you like seriously, except yourselves.
— Rudyard Kipling

Dedicated to the loving memory of Heiko Schlüter.
1942 – 2014

EXECUTIVE SUMMARY

Trust and trustworthiness facilitate interactions between human beings worldwide, every day. They enable the formation of friendships, making of profits and the adoption of new technologies, making life not only more pleasant, but furthering the societal development. Trust, for lack of a better word, is good. When human beings trust, they rely on the trusted party to be trustworthy, that is, *literally* worthy of the trust that is being placed in them. If it turns out that the trusted party is unworthy of the trust placed into it, the truster has misplaced its trust, has unwarrantedly relied and is liable to experience possibly unpleasant consequences. Human social evolution has equipped us with tools for determining another's trustworthiness through experience, cues and observations with which we aim to minimise the risk of misplacing our trust.

Social adaptation, however, is a slow process and the cues that are helpful in real, physical environments where we can observe and hear our interlocutors are less helpful in interactions that are conducted over data networks with other humans or computers, or even between two computers. This presents a challenge in a world where the virtual and the physical intermesh increasingly. A challenge that *computational trust models* seek to address by applying computational evidence-based methods to estimate trustworthiness.

In this thesis, the state-of-the-art in evidence-based trust models is extended and improved upon – in particular with regard to their statistical modelling. The statistics behind (Bayesian) trustworthiness estimation will receive special attention, their extension bringing about improvements in trustworthiness estimation that encompass the following aspects: *(i.)* statistically well-founded estimators for binomial and multinomial models of trust that can accurately estimate the trustworthiness of another party and those that can express the inherent uncertainty of the trustworthiness estimate in a statistically meaningful way, *(ii.)* better integration of recommendations by third parties using advanced methods for determining the reliability of the received recommendations, *(iii.)* improved responsiveness to changes in the behaviour of trusted parties, and *(iv.)* increasing the generalisability of trust-relevant information over a set of trusted parties. Novel estimators, methods for combining recommendations and other trust-relevant information, change detectors, as well as a mapping for integrating stereotype-based trustworthiness estimates, are bundled in an improved Bayesian trust model, *Multinomial CertainTrust*.

CONTRIBUTIONS Specific scientific contributions are structured into three distinct categories:

1. *A Model for Trustworthiness Estimation*: The statistics of trustworthiness estimation are investigated to design fully multinomial trustworthiness estimation model. Leveraging the assumptions behind the Bayesian estimation of binomial and multinomial proportions, accurate trustworthiness and certainty estimators are presented, and the integration of subjectivity via informed and non-informed Bayesian priors is discussed.
2. *Methods for Trustworthiness Information Processing*: Methods for facilitating trust propagation and accounting for concept drift in the behaviour of the trusted parties are introduced. All methods are applicable, by design, to both the binomial case and the multinomial case of trustworthiness estimation.
3. *Further extension for trustworthiness estimation*: Two methods for addressing the potential lack of direct experiences with new trustee in feedback-based trust models are presented. For one, the dedicated modelling of particular roles and the trust delegation between them is shown to be principally possible as an extension to existing feedback-based trust models. For another, a more general approach for feature-based generalisation using model-free, supervised machine-learners, is introduced.

EVALUATION The general properties of the trustworthiness and certainty estimators are derived formally from the basic assumptions underlying binomial and multinomial estimation problems, harnessing fundamentals of Bayesian statistics. Desired properties for the introduced certainty estimators, first postulated by Wang & Singh, are shown to hold through formal argument. The general soundness and applicability of the proposed certainty estimators is founded on the statistical properties of interval estimation techniques discussed in the related statistics work and formally and rigorously shown there.

The core estimation system and additional methods, in their entirety constituting the *Multinomial CertainTrust* model, are implemented in *R*, along with competing methods from the related work, specifically for determining recommender trustworthiness and coping with changing behaviour through ageing. The performance of the novel methods introduced in this thesis was tested against established methods from the related work in simulations.

Methods for hardcoding indicators of trustworthiness were implemented within a multi-agent framework and shown to be functional in agent-base simulation. Furthermore, supervised machine-learners were tested for their applicability by collecting a real-world data set of reputation data from a hotel booking site and evaluating their capabilities against this data set. The hotel data set exhibits properties, such as a high imbalance in the ratings, that appears typical of data that is generated from reputation systems, as these are also present in other data sets.

ZUSAMMENFASSUNG

Vertrauen und Vertrauenswürdigkeit erleichtern Menschen täglich und weltweit das Zusammenleben. Durch Vertrauen und Vertrauenswürdigkeit werden Freundschaften erst ermöglicht, Geschäftsgewinne erwirtschaftet und neue Technologien angenommen – in der Gesamtsicht wird durch Vertrauen nicht nur das Leben für jeden einzelnen angenehmer, sondern auch die Gesellschaft als Ganzes gestärkt. Vertrauen ist, einfach ausgedrückt, etwas gutes. Wenn man nun jemandem Vertrauen entgegenbringt, verlässt man sich darauf, dass die Seite, die das Vertrauen erhält auch vertrauenswürdig ist, d.h. sich buchstäblich des entgegengebrachten Vertrauens als würdig erweist. Sollte dies nicht der Fall sein, so hat der Vertrauende mit den unangenehmen Konsequenzen zurechtzukommen, die ungerechtfertigtes Vertrauen bedingt. Um sich vor solchen unwillkommenen und möglicherweise schmerzhaften Konsequenzen zu schützen, haben uns soziale Evolutionsprozesse mit Werkzeugen ausgestattet, um die Vertrauenswürdigkeit eines anderen zu beurteilen; durch gesammelte Erfahrungen, Beobachtungen des Auftretens des Gegenübers und subtil wahrgenommene Signale wird versucht, das Risiko des unbegründeten Vertrauens zu minimieren.

Dieser soziale Prozess und die erlernten sozialen Werkzeuge sind hilfreich im alltäglichen Leben und in der gewohnten, physischen Umgebung, in der wir Gesprächspartner beobachten können. Weniger hilfreich sind sie dort, wo uns die gewohnten Signale fehlen, besonders wenn Interaktionen in elektronischen Netzen stattfinden und unter Umständen das Gegenüber gar kein Mensch ist – oder gar die Kommunikation nur zwischen Maschinen stattfindet. In einer Umgebung, in der das Virtuelle und das Reale sich immer mehr miteinander verbinden, stellte die Abschätzung von Vertrauenswürdigkeit eine wissenschaftliche Herausforderung dar. Hierzu entwickelte Vertrauensmodelle nutzen datenbasierte Schätzmethoden, um Vertrauenswürdigkeit rechnerisch zu bestimmen,

Im Rahmen dieser Arbeit werden solche Vertrauensmodelle erweitert und verbessert. Hierbei liegt ein besonderes Augenmerk auf deren statistischer Methodik. Indem die (Bayes'sche) Statistik, die in solchen Modellen der Vertrauenswürdigkeitsabschätzung zugrunde liegt, betrachtet und für Erweiterungen genutzt wird, werden Verbesserungen bezüglich folgender Aspekte erreicht: *(i.)* statistisch wohl-begründete Schätzer zur Bestimmung von Vertrauenswürdigkeit und der bei der Schätzung auftretenden, inhärenten Unsicherheit sowohl für binomial, als auch für multinomiale Modelle, *(ii.)* eine Verbesserung der Integration von Empfehlungen durch dritte, die insbesondere durch fortgeschrittene Methoden zur Bestimmung der Zuverlässigkeit der

Empfehlenden ermöglicht wird, *(iii.)* Verbesserung der Empfindlichkeit bei der Entdeckung von Verhaltensänderungen im Zusammenhang mit vertrauenswürdigem Verhalten und *(iv.)* Verbesserung der Generalisierbarkeit von Daten, die für die Vertrauenswürdigkeitsabschätzung herangezogen werden können.

Die entwickelten neuartigen Schätzer, Methoden zur Kombination von vertrauensrelevanten Informationen, Detektoren für Verhaltensänderungen, genauso wie eine Schnittstelle zur Integration von Vertrauenswürdigkeitsabschätzungen anhand von Stereotypen aus maschinellem Lernen, sind integriert in ein fortschrittliches Bayes'sches Vertrauensmodell, genannt *Mutlinomial CertainTrust*.

BEITRÄGE Die erbrachten Beiträge gliedern sich in drei Kategorien:

1. *Ein Modell für Vertrauenswürdigkeitsabschätzung*: Die statistischen Grundlagen der Vertrauenswürdigkeitsanschätzung werden untersucht, um darauf aufbauend ein vollständig multinomiales Vorhersagemodell für Vertrauenswürdigkeit zu entwickeln. Hierzu werden, ausgehend von den Grundannahmen Bayes'scher Schätzverfahren für binomial und multinomiale Verhältnisse, genaue Schätzer für Vertrauenswürdigkeit und Unsicherheit vorgestellt; zudem wird die Integration von subjektiven Information über informative und nicht-informative Bayes'sche A-priori-Wahrscheinlichkeiten diskutiert.
2. *Methoden zur Verarbeitung von vertrauensrelevanten Informationen*: Zur Verbesserung der Propagation von Vertrauen und der Detektion von Verhaltensänderungen werden entsprechende Methoden dem Vertrauensmodell hinzugefügt. Alle Methoden sind hierbei sowohl für binomial als auch für multinomial Vertrauenswürdigkeitsabschätzung anwendbar.
3. *Weitere Methoden zur Vertrauenswürdigkeitsbestimmung*: Zwei Ansätze, die den etwaigen Mangel an direkten Erfahrungen mit einzelnen Interaktionspartnern addressieren, werden vorgestellt. Zum einen wird das dedizierte Abbilden von Rollen innerhalb von erweiterten Vertrauensbeziehungen und die Delegation zwischen diesen als prinzipielle Möglichkeit gezeigt. Zum anderen wird eine generellerer Ansatz durch die Anwendung von überwachtem maschinellen Lernen auf Basis von modellfreien Lernverfahren eingeführt.

EVALUATION Die Eigenschaften der Schätzer für Vertrauenswürdigkeit und Unsicherheit werden formal anhand der statistischen Eigenschaften des Bayes'schen Schätzverfahrens gezeigt. Gewünschte Eigenschaften für Unsicherheitsschätzer, eingeführt vorn Wang & Singh, werden anhand von formaler Argumentation nachgewiesen. Durch die Zurückführung der Schätzverfahren auf etablierte grundlegende statistische Modelle von Punkt- und Intervalschätzern sind die generellen Eigenschaften der Schätzer durch die verwandte Literatur im Bereich Statistik und Schätztheorie gestützt

Das Modell zur Vertrauenswürdigkeitsabschätzung als Basis, sowie zusammen mit den weiteren entwickelten Methoden, wurden in der

Programmiersprache *R* implementiert, ebenso wie Methoden aus der Literatur. Die in dieser Arbeit eingeführten Verfahren, insbesondere zur Bestimmung der Zuverlässigkeit von Empfehlungen und der Detektion von Verhaltensänderungen wurden gegen Methoden aus der verwandten Literatur getestet und mit diesen verglichen.

Die Methoden zur Vertrauenswürdigkeitsgeneralisierbarkeit mittels Rollen wurde in einer Multiagentensimulation implementiert und getestet. Die Anwendbarkeit von maschinellem Lernen wurde anhand eine Datensatzes, der aus einem Reputationssytem für Hotelbuchungen gewonnen wurde, evaluiert. Der verwendete Datensatz zeigte Eigenschaften, die als repräsentativ für Datensätze betrachtet werden können, wie sie von Vertrauens- und Reputationssystäemen typischerweise generiert werden.

Without trust there's no friendship, no closeness.
None of the emotional bonds that makes us who we are.

— Commander William T. Riker,
Star Trek: The Next Generation, "Legacy" (1990)

ACKNOWLEDGMENTS

This thesis would not have come into existence without the help and encouragement of colleagues, friends and family. My gratitude and thanks go out to all of them.

First, warmest thanks to my advisor Max Mühlhäuser, for his continued support, confidence and advise in matters of research and beyond. Second, I am grateful to Christian Jensen for acting as a second referee.

A big 'thank you' also to all my current and former colleagues at the Telecooperation Group of TU Darmstadt, who have made a home out of a place of work and have become friends, especially all of those at the areas Smart Security and Trust and Smart Secure Infrastructures Jörg Daubert, Mathias Fischer, Carlos Garcia, Sheikh Mahbub Habib, Shankar Karuppayah, Leonardo Martucci, Sebastian Ries, Stefan Schiffner, Emmanouil Vasilomanolakis, Florian Volk. They, and everybody else at TK, made working a pleasure! The same goes for the great staff at the group: Elke Halla, Nina Jäget, Silke Romero, Karin Tillack – thank you for making work at TK great!

Working on projects has been a big part of the experience at TU Darmstadt. Here, too, many people have made this experience richer: Sebastian Biedermann, Michael Riecker, Rachid El Bansarkhani, to name just a few.

I am grateful to all my friends for their support over the years. Thank you for believing in me! Dominik Heider, Martin Pyka, Robert Bräuning and Markus Borschbach have not only been steadfast friends but have also contributed to this thesis with comments and discussion.

Finally, thanks to my family. Without you, doing this would have been impossible. I love you!

CONTENTS

LIST OF TABLES

1 INTRODUCTION

Trust is an important concept, encountered and leveraged in everyday life to guide decisions that have to be made in under risk and uncertainty. Its application in social life comes intuitively, stemming from a long evolutionary process. Its application in digital life, however, requires explicit modelling and adjustment of familiar trust techniques.

1.1 TRUST IN SOCIAL LIFE

Kenneth J. Arrow, the youngest ever Nobel laureate in economics, remarked on the importance of trust that

> *trust is an important lubricant of a social system. It is extremely efficient; it saves a lot of trouble to have a fair degree of reliance on other people's word* [7].

In other words, when trust flourishes and is rewarded, societies prosper; when distrust and egoism take hold, the lubricant that enables the cog wheels of societal interaction to smoothly turn and interlink dries up – and societies decline (see Putnam [164]).

From a macroscopic point of view, the positive effects of trust on societies are indisputable. The macroscopic benefits of trust emerge from a more microscopic scale, that of interaction between people within society. A common understanding of trust seems to be innate in human beings, and a such a common understanding can be put into words by consulting popular dictionaries. The Merriam-Webster Online Dictionary, for instance, defines trust as the

> *assured reliance on the character, ability, strength, or truth of someone or something*.

It further provides another definition that defines trust as a

> *dependence on something future or contingent: hope*.

These two definitions already encapsulate several key aspects of what are the core elements of trust:

- *dyadic, directed relation*, between two parties, marked by
- *dependence* of one party (the *truster*) on the other (the *trustee*), under
- *risk and uncertainty*, which is contingent on
- *internal qualities* of the trustee, supported by

- *assurances* of the goodness of the trustee's qualities.

These aspects can be considered constituent components of trust and the act of trusting at its most basic level. While the above definitions consider trust an act or an action, these components are also applicable to other interpretations of trust beyond its behavioural expression, including those discussed in Chapter 2.1.

When elaborating on the constituent components given above, it is helpful to frame the dependence of the truster on the trustee in terms of an *interaction*. Although one can trust in inanimate objects (e.g., that a piece of rope will be sufficient to support the truster's weight) or even constructs of the mind (e.g., that a particular statement is true), for the sake of illustration assume that truster and trustee are both active agents. The truster relies on the trustee to perform a particular task (or possibly to refrain from performing it), from which the truster hopes to derive some form of benefit once the interaction is concluded.

Using such an interaction as an illustrative basis, the relational quality of trust is obvious. The notion of a *dyadic* relation stems from the two roles involved, the truster and the trustee, while the *directedness* of the relation is due to the potential difference of the internal qualities of the truster and the trustee. In other words, because the *'... character, ability, strength, or truth ... '* are inherent qualities of the trustee, and hence are not necessarily identical for truster and trustee, the relation is *directed*. That is, the trust a specific truster, say A, extends towards a specific trustee, say B, is generally different when the roles are reversed. In other words, trust of A in B does not imply trust of B in A.

By relying on qualities of the trustee that are difficult to establish – at least a-priori – such as '*character, ability, strength, or truth*', the outcome of the interaction becomes contingent on the actions of the trustee, making the truster dependent of the trustee's actions or inactions. A situation of dependence on others, combined with uncertain outcomes is ripe with risks. The dangers of deception, of unreciprocated trust, simply put 'of being let down' are an integral part of trusting. Such an acceptance of risk should be informed, and the truster is well-advised to have at least an inkling of the trustee's internal qualities, that is, its ability and willingness to reciprocate the trust put into it. In the following, this willingness and ability to reciprocate the trust put into it is considered the trustee's *trustworthiness*.

However, where there is risk, there is also chance. Of course, when extending trust, the chance of a positive outcome should outweigh the risk of a negative one. The utility of an interaction, that is, the expected gain compared to the expected loss, is a natural decision criterion for trusting [44].

Given the difficulty of establishing the trustworthiness of the trustee, achieving *assured* instead of blind reliance requires a way of estimat-

ing the potentially unobservable internal qualities of the trustee. In social contexts, trusters can avail themselves of various sources of information in order to assure themselves of a trustee's trustworthiness , such as:

- *experience*: by accruing historical evidence of the past behaviour of the trustee and whether or not such behaviour was deserving of trust, the truster leverages a direct relationship with trustee for assurance. In a social context, this requires truster to become familiar with the trustee, a potentially time consuming process.

- *cues*: since the accumulation of experience is a time consuming process, humans have learnt, over the course of our evolution, to identify *social cues* [6] that are typically associated with trustees deserving of trust and that can be used to persuade [134] establish rapport [72]. Typical social cues can be verbal, such as tone of voice, or nonverbal, presenting in body language, posture, and gestures. Additionally, physical or emotional proximity [30] are also considered social cues.

  These cues, together with other readily identifiable features (such as tidiness of appearance or a recognisable affiliations) can be used to form *stereotypes* of expected trustee behaviour. Because of the generalisation present in the formation of stereotypes, trust built upon social stereotypes are prone to exploitation, in particular when supposed marks of what makes a 'good' trustee turn out to be non-representative.

- *recommendations*: historical evidence collected by others can be leveraged by the truster in the form of recommendations by trusted sources, such as family, friends or others, which the truster believes to be knowledgeable. Recommendations enable a limited form of *transitivity* for trust relations: truster A relies on trustee B based on a recommendation by trusted recommender C, relying on C's expertise in recommending.

Finally, trust is dependent both on specific *contexts* and *situations* (see, for example, [112]). Context can be considered a topical aspect of trust; one trusts someone else in a specific context, to which the trustee's ability are suited, while in other contexts, to which the trustee's abilities are not deemed suitable, trust is not extended. For instance, A may trust B when it comes to fixing a faucet, but not when it comes to assemble clockwork. Situation is an immediate aspect of trust; one might not generally trust someone else in a specific context under *normal condition*, but decide to trust in the same context under *abnormal conditions*. For instance [112], A might not trust a rope for rappelling from a window under normal conditions, but might in a situation where the house is ablaze with fire.

The preceding section illustrates that trust, even though it comes natural as a social tool and appears to have an accessible definition, is not a trivial concept and will be covered in more detail in Chapter 2.1. It is, however, highly useful and intuitively applicable by most people in the social interactions we are used to. In digital life, however, determining whether or not to trust is not supported by the (evolutionary) learned toolset we use in social interactions.

1.2 COMPUTATIONAL TRUST IN DIGITAL LIFE

Particularly in online environments, where traditional ways of establishing the trustworthiness of another party – that have been a staple of human interaction – cannot be readily applied [141], computational methods for estimating trustworthiness are useful tools (see generally [113]). The use of online networks in every day life, be it for business or for pleasure, has become ubiquitous. A few short years ago, access to the Internet was limited by access to both a desktop computer and a landline connection. Advances in wireless communication and miniaturisation have done away with these restrictions. Since 2000, the number of worldwide Internet users has increased from 394 million users to more than 2.9 *billion* users in 2014; the global Internet advertising revenue, an indicator of the commercial viability of the Internet, amounted to approximately 117 billion US dollars in 2013 and is projected to increase to more than 194 billion US dollars worldwide in 2018[1].

Online fraud is the ugly by-product of the flourishing of Internet commerce; in 2013, the FBI reported total combined losses of more than 780 million US dollars (a 48.8 per cent increase from 2012) from more than 262,000 reported cases of online fraud [95] for the United States. The number of unreported cases may be considerably higher; the responsible official at the FBI, the managing director of the white collar crime centre, John Kane, estimated in 2009 that as few as 15 per cent of cases are reported [125]. Most of the fraudulent behaviour reported in [95] exploited social interactions. Thus, primarily technical solutions, such as integrity checks based on a trusted platform architecture, alone are insufficient to protect users. Trust, as a soft security mechanism, therefore serves an important role in deciding whether or not to interact with another party. Systems that determine the trustworthiness of other empower the user to make informed decisions and form collaborations that turn out to be successful.

The success of collaboration depends on the successful selection of reliable and trustworthy partners in a techno-social environment. Relying only on traditional certificate-based approaches alone is insufficient. First, in an unmanaged environment, there is no completely trusted central authority. Such an authority would be necessary as

1 Source: www.statista.com

an anchor for issuing and revoking certificates that themselves are fully trusted. Attacks on certificate authorities, for instance *DigiNotar*, *Comodo* or *RSA*, have underscored the vulnerability of a certificate-only approach. Second, certificates normally only provide information about the identity of an entity. This is also insufficient, as a mere identifier or pseudonym does not convey information about the behaviour of an entity. Rather, compliant behaviour would have to be enforced by the certificate authority or some other third party that needs to be trustworthy and fully trusted.

Estimating trustworthiness in a traditional way, e.g., from social cues, is difficult in electronically-mediated interactions on the Internet. However, the collection of feedback, data and statistics is comparatively easy – a fact that computational trust and reputation systems leverage in order to provide a (probabilistic) estimate of another parties future behaviour.

Computational methods for determining trust can not only secure the trusting party against loss, but actually drive the adoption of new technologies and services. New technologies are – by the very virtue of being new – unproven and unfamiliar to a potential user. Its qualities might not yet be entirely obvious and its benefit an as of yet unsubstantiated promise. Particularly new digital technologies, that lack an physical embodiment, make it hard for a potential user to assess their quality and build trust. Computational trust mechanisms can be used explicate the quality of a new technology, create an environment that is accountable and assist potential users to make an informed decision whether or not to adopt a new technology.

This thesis addresses the interesting challenge of developing and extending methods for trustworthiness estimation that are statistically meaningful and provide the trusting parties with tools to reliably assess the trustworthiness of those they intend to depend upon. Trustworthiness estimation is at the core of probabilistic computational trust models. A trustworthiness estimate, also referred to as a *trust score*, is a probabilistic score representing a truster's assessment that a trustee will act in a certain way. In order to compute a trust score, computational trust models provide means for processing evidence, for example from past experience and recommendations, and for using this evidence to compute a reliable estimate of a trustee's future behaviour.

Informally speaking, a very simple application of computational trust can be described as follows: say A wants to determine the trustworthiness of a potentially trustee B using computational trust methods. For this, A recalls its past interactions with B, if any, and asks acquaintances and experts it knows for their recommendations with regard to B. A combines its own knowledge from past interactions with B with the recommendations it has received, accounting for how reliable A believes each recommender to be. From this base of ev-

idence, A then determines how likely B is to act in a way that A deems satisfactory, using an appropriate estimator. This results in a probabilistic trust score, based upon which A decides whether or not to trust and interact with B. If the decision is a positive one, A interacts with B, gaining more experience with B that can be used to estimate B's trustworthiness more accurately in the future.

1.3 GOAL AND OBJECTIVE OF RESEARCH

The goal of this thesis is to provide improved trustworthiness estimation techniques for probabilistic computational trust models. In environments that are not centrally managed by a completely trusted and trustworthy third party, such models have to allow each user individually to collect data and compute meaningful trustworthiness estimates on its (potential) interaction partners. This is a subjective, distributed and largely sampling-based statistical procedure. Under these circumstances, estimates are generally affected by a variable degree of uncertainty and information maybe scarce, at least locally. Consequently, estimation procedures do not only have to be accurate and statistically sound, the also have to provide for computing and conveying the inherent uncertainty of an estimate. Furthermore, procedures for processing the information that is available for making an estimate are required to overcome data scarcity – ranging from the summation and aggregation of data to more sophisticated procedures that possibly allow for the generation of new cues or stereotypes, in order to provide a way of generalising trust information.

The objective of research for this thesis is the realisation of these goals in the form of an expressive trust model, building upon established statistical methods and previous work in the field of trust modelling. At its core, this requires the establishment of a well-defined trustworthiness estimation model that is paired with a sound uncertainty estimator. Such an estimation model should be based on evidence, either from direct experiences that were made in the past, or from reported experiences in the form of recommendations from others. The adaptation of point and interval estimation methods from the statistics literature to work, in conjunction with existing trust models, serves as a starting point.

In order to process recommendations from others and enable (limited) transitivity for trust, methods for the combination of one's own and other's experiences have to be provided, that allow for the robust integration of both kinds of experiences. For this, the reliability of recommendations from others has to be established and appropriate weighting procedures have to be introduced. Additionally, behaviour can change over time, sometimes fundamentally. This too has to be detected and addressed.

When developing adapted and novel methods, the focus is on their statistical basis. Issues range from interval-based estimation as a foundation for certainty estimation, to hypothesis testing in determining the reliability of recommendations and detecting changes in behaviour. Providing this for both binary *and* more generally for any kind of m-categorical evidence, that is, for both binomial and multinomial models of trust, provides a further challenge.

1.4 SCIENTIFIC CONTRIBUTION AND EVALUATION

This thesis proposes novel methods for trustworthiness estimation, realised within the framework of a completely overhauled version of the binomial *CertainTrust* model [173], resulting in the *Multinomial CertainTrust* model. The approach is based on estimation-theory and Bayesian statistics to provide a statistically solid core system for trustworthiness estimation with both binary and m-categorical, $m > 2$, feedback. This core is then extended with

1.4.1 *Contributions*

1. **An extended model for binomial and multinomial trustworthiness estimation:** In the present thesis, the statistics behind the *CertainTrust* model [173] are reformulated, motivated, revised and extended. In particular, the certainty estimation is given a new interpretation, formally derived from the binomial and categorical distributions underlying the binomial and multinomial case, respectively. This interpretation considers certainty an estimate of the dispersion of the trust score computed in *CertainTrust*. Fundamentally, this also advances the state-of-the-art presented in works by Wang & Singh [196] and Teacy et al. [189], leverages proven statistical methods (see, e.g., [27]), and provides a flexible extension from the binomial into the multinomial case of trust assessment.

   The binomial *CertainTrust* model is extended to the *Multinomial CertainTrust* model, providing simultaneous confidence interval-based certainty estimators and graphical representations of the results.

   Specific contributions include:

   - For the *binomial case* of trustworthiness assessment:
     - *Credibility Interval-based Certainty Estimator*: A dispersion-based certainty estimator, derived from the Bayesian Jeffreys credibility interval for binomial proportions.
     - *Confidence Interval-based Certainty Estimator*: A dispersion-based certainty estimator, derived from the frequentist Wilson confidence interval for binomial proportions;

providing a closed-form alternative to the open-form Credibility Interval-based Certainty Estimator at comparable performance levels.

  - An adjusted computation of the *CertainTrust* expectation value, in order to incorporate the novel certainty estimators into the *CertainTrust* model (Section 3.1.7).
  - An augmented human trust interface (HTI) capable of displaying the potential dispersion of a trust estimate.

- For the *multinomial case* of trustworthiness assessment:
  - *Multinomial CertainTrust*: A complete extension of the predictive model behind *CertainTrust* to handle multinomial opinions
  - *Simultaneous Credibility Interval-based Certainty Estimator for Multinomial Proportions*: A version of the Credibility Interval-based Certainty Estimator that corrects for the multiple testing inherent in multinomial proportions.
  - *Simultaneuous Confidence Interval-based Certainty Estimator for Multinomial Proportions*: A closed-form alternative to the Simultaneous Credibility Interval-based Certainty Estimator, using Goodman's correction of the Wilson confidence interval.
  - A mapping of multinomial priors to *Multinomial CertainTrust* initial trust parameters and corresponding *MultinomialCertainTrust* expectation value computation.

Overall, a complete prediction model for binomial *and* multinomial trustworthiness assessment is provided, representing the core of a more comprehensive trust model.

2. **Extended methods for trust propagation, combining evidence and coping with changes in behaviour:** The core estimation model is augmented with further methods for facilitating trust propagation and accounting for concept drift in the behaviour of the trusted parties. All of these augmentations are applicable, by design, to both the binomial case and the multinomial case of trustworthiness estimation.

   Specifically, these methods deal with determining recommender trustworthiness and the combination of opinions; both of which are necessities for robust trust propagation. Additionally, non-stationarity in the data generating process, i.e., potentially changing and dynamic trustee behaviour, is introduced to the model assumptions.

   The methods for combining evidence encompass discounting, consensus and fusion operations. The operations for discount-

ing and consensus were modified from their original form as formulated in [103, 173] to fit the extended version of the *CertainTrust* model introduced in this thesis, *Multinomial CertainTrust*. This represents a necessary step in providing a comprehensive trust model that includes capabilities for trust propagation. Additionally, the fusion operation, a method for combining opinions for which the assumption of independence does not hold, is adapted for use in *Multinomial CertainTrust*. The original version, which is essentially an averaging operation present in both *Subjective Logic* [104] and *CertainLogic* [77, 175], has been adapted to *Multinomial CertainTrust*. Weighted and conflict-aware extensions, first published in our prior work [77], have been presented and considerably extended to the multinomial case, including a novel method for computing the degree of conflict leveraging the an exact hypothesis test (*Fisher's Exact Test* (*FET*)). Together with the novel *FET*-based method for determining recommender trustworthiness, a comprehensive multinomial trust model with trust propagation capabilities is enabled.

Furthermore, dynamicity, in the form of non-stationary behaviour by a trusted party, is addressed and handled by applying state-of-the-art change point detection to trustworthiness estimation. Compared to multiplicative ageing, the application of change point detection method does not affect the achievable accuracy of the trust estimator. When used in conjunction with ageing, change point detection improves the responsiveness of the estimator to behavioural change over either individual method.

Specific contributions include:

- A novel estimation method for *recommender trustworthiness estimation* is introduced that compares favourably to the related work. This *FET*-based recommender trustworthiness estimation method can be applied to multinomial evidence, as opposed to the methods from the related work, which are applicable to binomial models only.
- Operations for trust propagation are adapted or newly introduced for the use with *Mutlinomial CertainTrust*, specifically:
  - *Discounting* is adapted to *Mutlinomial CertainTrust* opinions,
  - *Consenus* is adapted to *Mutlinomial CertainTrust* opinions,
  - *Average Fusion* is adapted to *Mutlinomial CertainTrust* opinions,

  - *Weigthed and Conflict-aware Fusion* are newly introduced and expanded from our own prior work [77], including a novel way of computing the degree of conflict between opinions.
- *Change point detection* is introduced into trustworthiness estimation in order to improve the trust model's responsiveness to dynamicity, expressed as non-stationarity.

Overall, extensions necessary to make the core trustworthiness estimation model a comprehensive trust model with trust propagation capabilities are provided.

3. **Further extension for trustworthiness estimation:** Two methods for addressing the potential lack of direct experiences with new trustee in feedback-based trust models are presented. For one, the dedicated modelling of particular roles and the trust delegation between them is shown to be principally possible as an extension to existing feedback-based trust models. For another, a more general approach for feature-based trustworthiness estimation using model-free, supervised machine-learners, is introduced.

   Generally speaking, the primary goal of the methods introduced and applied in this chapter is to imbue feedback-based trust models with the ability to determine trust for individual, unproven trustees. That is, to allow an estimation of trustworthiness of a trustee based on features and connections to specific other parties said trustee exhibits and that can be observed by a truster. This is partially derived from the social practice of learning *stereotypes* and (pre-)judging or discriminating according to these.

   Specific contributions include:

   - Trust model extensions to provide trust-relevant information by leveraging specific roles and relations that can be encountered in e-commerce interactions. Three specific examples are chosen in this thesis to show the principal practicability of such extensions:
     - *certifiers*, which certify the service quality, and hence the trustworthiness, of a certified trustee. A trust delegation mechanism is provided for partially transferring trust in a certifier onto the certified trustee.
     - *insurers*, which provide assurance against loss potentially incurred from an untrustworthy trustee. A trust delegation mechanism is provided that influences decision trust, that is, the expected utility of an interaction with the insured trustee.

  - *coalition partners*, which are associated with the trustee in a (semi-)permanent fashion. A trust delegation mechanism is provided for partially transferring trust in coalition partners onto the trustee partner in a coalition.

- Application and evaluation of powerful supervised machine learning approaches to a real-world data set with a regressand value generated from a reputation system. The distribution of the regressand value follows a distribution that is both typical of those from a reputation system and is adverse to the successful application of supervised methods. The predictive results suggest that the model-centric approach taken in the design of existing stereotyping trust models needs to be complemented by a data-centric analysis and that idealised simulations are insufficient to ascertain feasibility.
- A mapping from the output of supervised machine learners to *CertainTrust* opinions, enabling the integration of supervised learning with feedback-based trust models. Thereby, generalisable information contained in the features of a given data set can be harnessed, even if the prediction quality is only mediocre. The prediction of the estimator is mapped directly to the *CertainTrust* trust parameter t, while a statistical measure of the prediction quality – in this case, the *normalised root mean squared error* (NRMSE) – is mapped to the certainty parameter c.

Overall, methods for improving trustworthiness estimates in feedback-based trust models under scarce information for individual trustees are proposed. First indicators of trustworthiness were hardcoded into an existing probabilistic trust model. Second, an approach to flexibly include stereotype-like results from non-parametric, model-free supervised learners was used to extend feedback-based trust models. Particularly the results gathered from the application of the latter points at a further need for researching trust models not just from a model-centric, but rather also from a data-centric point of view.

1.4.2 *Evaluation*

The general properties of the trustworthiness and certainty estimators are derived formally from the basic assumptions underlying binomial and multinomial estimation problems, harnessing fundamentals of Bayesian statistics. Desired properties for the introduced certainty estimators, first postulated by Wang & Singh [196], are shown to hold through formal argument. The general soundness and applicability of

the proposed certainty estimators is founded on the statistical properties of interval estimation techniques discussed in the related statistics work, particularly [27] and formally and rigorously shown there.

The core estimation system and additional methods, in their entirety constituting the *Multinomial CertainTrust* model, are implemented in *R*, along with competing methods from the related work, specifically for determining recommender trustworthiness and coping with changing behaviour through ageing. The performance of the novel methods introduced in this thesis was tested against established methods from the related work in simulations.

Methods for hardcoding indicators of trustworthiness were implemented within a multi-agent framework [45, 82] and shown to be functional in agent-base simulation. Furthermore, supervised machine-learners were tested for their applicability by collecting a real-world data set of reputation data from a hotel booking site and evaluating their capabilities against this data set. The hotel data set exhibits properties, such as a high imbalance in the ratings, that appears typical of data that is generated from reputation systems, as these are also present in other data sets.

1.5 PUBLICATIONS

Some parts of this thesis builds on work that has been published before in conferences and journals. Some of the related work on trust models builds on [76]. The methods for adding generalisability to trust models were previously published; a paper on hardcoding indicators of trustworthiness appeared in [84], while the application of supervised machine learning to trustworthiness estimation was discussed in [85, 86]. The agent-based simulation framework used in Chapter 5 was originally developed for [82]. Some of the author's work on trust propagation, in a wider context, inspired parts of this thesis [80, 81, 83].

1.6 THESIS STRUCTURE

The remainder of this thesis is structured as follows. Chapter 2 provides background information about the concept of trust, presents assumptions and requirements, as well as related work on trust models and the statistics literature mainly used in this thesis.

Chapters 3, 4, and 5 contain the main contributions of this thesis, as listed in Section 1.4.1. Chapter 3 provides an extended model for binomial and multinomial trustworthiness estimation, building on *CertainTrust*. The binomial estimation model behind *CertainTrust* is extended in two directions: first, its certainty estimation component is replaced with a statistically well-founded estimation mechanism harnessing interval estimation techniques; second, the binomial estimation model

is extended to a multinomial estimation model, facilitating fine granular feedback categories. The result of this second extension, *Multinomial CertainTrust*, also features interval-based certainty estimation.

Chapter 4 provides the extensions necessary to build a sophisticated trust system from the trustworthiness estimation model introduced in Chapter 3. This includes methods for weighting and combining evidence via discounting, consensus and fusion operations, adapted for application within *Multinomial CertainTrust*. Exact hypothesis testing is introduced as a novel method to provide similarity measures when determining recommender trustworthiness, the degree-of-conflict in conflict-aware fusion and in detecting changes in behaviour.

Chapter 5 proposes two mechanisms as extensions to feedback-based trust models. First, the integration of so-called indicators of trustworthiness is discussed and shown via three distinct examples of what could constitute such an indicator. Then, a more general method using supervised machine learning is discussed and tested against a real-world reputation data set.

Finally, Chapter 6, concludes this thesis and provides an outlook.

2

BACKGROUND AND RELATED WORK

In this chapter, the necessary background information with regard to this thesis is introduced and the corresponding related work is briefly discussed. First (Section 2.1), related work on trust and trustworthiness is introduced, very briefly outlining the derivation of the probabilistic, computational notion of trust from its social origins. Then (Section 2.2), existing computational trust models are presented, including general assumptions made when considering trustworthiness estimation within this thesis. From these assumptions, a number of requirements are derived and the most closely related trust models are checked against them. Finally (Section 2.3), further estimation theoretic and statistical methods that will be relevant in the latter part of this thesis will be given.

2.1 CONCEPTS OF TRUST AND TRUSTWORTHINESS

The research into trust is very much influenced by its nature as a *social* concept. Hence, the social sciences have a long tradition of research into the field of trust. This research has produced a rich landscape of literature on the subject, yet little consensus on the specific meaning of trust has been reached [142]. This may well be due to the universal importance of trust-related concepts in a multitude of socially relevant conditions. As such, trust enables and lubricates cooperative endeavours [7, 44, 64], permits positive interpersonal relationships [57, 127], determines how we interact among each other [14, 69], reduces uncertainty [146, 200, 201], or determines the effectiveness of working relationships [62]. Thus, while trust and its indisputably positive effects on social interactions have been widely recognised, the very diversity of scholarly application fields and investigating (social science) research disciplines influence the perception of what trust is and how it should be defined. Trust itself has been described as an "elusive concept" [64, 144, 205], with the myriad of trust definitions forming a "conceptual morass" [10, 34]. Some trust definitions from the literature are listed in Appendix A, p. 227.

In most definitions of trust listed in Appendix A, the notion of dependence in some kind of interaction is expounded. In order for trust to be a useful concept, one party has to be willing to depend on the other, without being able to actively control the other party's actions. In the parlance of the trust literature, the depending party is referred to as the *truster*, while the party that is being relied upon is being referred to as the *trustee*. The term *interaction*, as used here,

is used in a slightly more abstract manner than in its everyday usage. In the following, the term interaction encapsulates a situation of dependence between truster and trustee that can be evaluated by the truster according to its outcome. As such, an interaction does not necessarily require an action on the part of the trustee. Nor does it require the truster and the trustee to actually interact with each other in the conventional sense. As an example for the former, an interaction can involve the truster trusting the trustee *not* to act in a particular way, for instance not to divulge information to a third party. As an example for the latter, a truster may trust the trustee to be present at a certain place at a certain time, without explicitly informing or reminding the trustee. Both would be considered interactions in the following.

In the following, the work by McKnight et al. on the dimensions of the high level concept of trust will be briefly introduced, from which a working definition of trust for this thesis will be derived.

2.1.1 *Differentiating Trust*

Differentiating, defining and classifying different dimensions of the high level concept of trust as used in the literature is a necessary step in understanding the fundamental processes of trust establishment. McKnight et al. [142, 143, 144, 145] have expended considerable effort in reviewing and aggregating views of trust from diverse fields of study, resulting in an influential conceptual typology of the components involved in the formation of trust-based relations [145]. Accordingly, trust can be decomposed into five fundamental categories that nonetheless influence each other, as depicted in Figure 1. In the following, these categories are briefly reproduced from [145].

TRUST-RELATED BEHAVIOUR is the most manifest representation of trust. As such, *trust as a behaviour* has been a prominent conceptualisation in the literature [44, 57]. It is defined as "*the extent to which one person voluntarily depends on another in a specific situation with a feeling of relative security, even though negative consequences are possible*" [142]. The behavioural aspect effects a transformation from willingness to depend into actual dependence, essentially allowing the trustee to gain power over the truster, from which an assumption of risk on the part of the truster ensues.

TRUSTING INTENTIONS are closely correlated to trust-related behaviour. While trust-related behaviour represents the actual action of depending on the trustee, trusting intentions means the truster's secure *willingness* to depend on the trustee. McKnight et al. [145] distinguish two sub-constructs of trusting intentions:

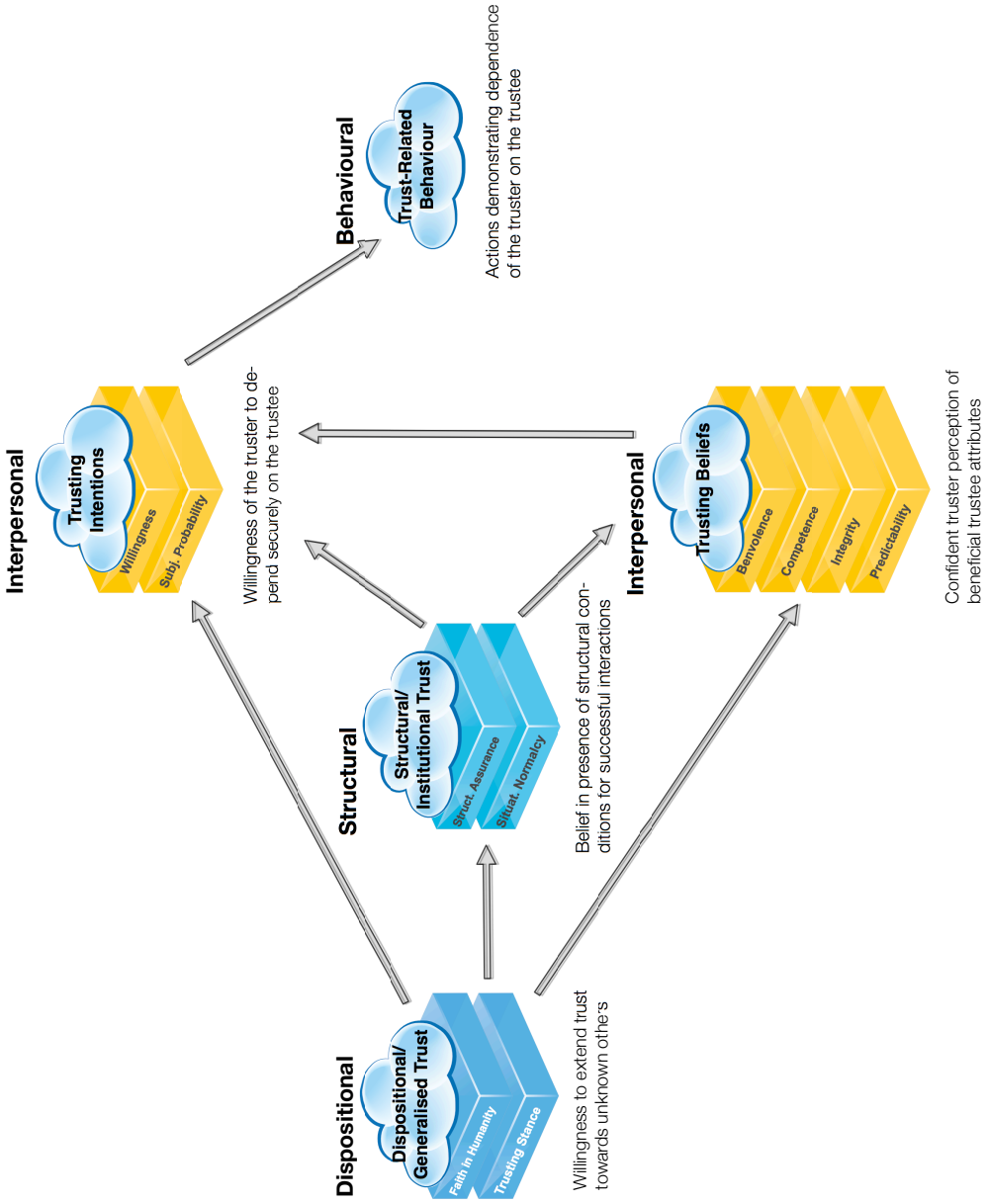


Figure 1: Typology of trust according to McKnight et al. [143, 144]

- *Willingness to Depend* – the general, voluntary preparedness of the truster to make him-/herself vulnerable to actions of the trustee,
- *Subjective Probability of Depending* – the likelihood, as perceived by the truster, to make him-/herself dependent on the trustee in specific ways.

While a truster's willingness to depend on a trustee can be expressed by agreeing to general statements about relying on the trustee, the subjective probability of depending is *situation-dependent* [145]. Thus, you might be willing to trust your neighbour to clear the sidewalk of snow in the winter, but not willing to trust him to look after your children.

TRUSTING BELIEFS refer to the confident truster perception that the trustee has certain attributes that benefit the truster. This comprises those facets of the internal representation the truster has constructed of the trustee that are relevant to a successful interaction between them. McKnight et al. [145, 144] have identified four specific beliefs that feature prominently in this process, namely

- *Competence Belief* – the truster's belief that the trustee is able to fulfil its obligations,
- *Benevolence Belief* – the truster's belief that the trustee has no ulterior motives,
- *Integrity Belief* – the truster's belief that the trustee will keep its commitments and does not lie,
- *Predictability Belief* – the truster's belief that the trustee's future actions are in accordance with his/her prior (observed) behavior.

The truster's belief that the trustee possesses these traits has a direct impact on his/her trusting intentions. If, for instance, the truster perceives the trustee to be lacking the ability to perform a particular task, its intention to delegate that task to the trustee will be low and the interaction between truster and trustee will be abortive.

INSTITUTION-BASED TRUST is the belief that the environment in which a particular interaction takes provides the required structural conditions that facilitate a successful interaction between truster and trustee. These structural conditions include technological (e.g. reliable data encryption), legal (e.g. enforceable laws governing interactions) and social safeguards (e.g. ostracising of unreliable community members). McKnight et al. [145, 144] define two dimensions of institution-based trust, *structural assurance* and *situational normality*.

While structural assurance means the belief that structural safeguards are in place (e.g. data encryption and legal asset protection in an online banking scenario), situational normality refers to the belief that a particular interactions is conducted under conditions that are representative of the environment and are therefore favorable to a successful interaction.

DISPOSITION TO TRUST is the most general category of trust presented by McKnight et al. [142, 144]. It is both independent of other *specific* interactors and the situational context. Its two sub-constructs, *faith in humanity* and *trusting stance*, deal with generalised views of indiscriminate groups of persons. Faith in humanity describes the general extent to which the trusters beliefs others to be competent, benevolent and morally upright, irrespective of specific personal traits. The trusting stance represents a personal approach to interacting; rather, "*regardless of what one beliefs about peoples' attributes, one assumes better outcomes result from dealing with people as though they are well meaning and reliable*" [144, 176].

The aforementioned types of trust – dispositional, institutional, interpersonal and behavioural – influence each other not only in the process of deriving a decision whether or not to interact with another entity (the feedforward direction), but also through the experiences made with specific others after an interaction has occurred (the feedback direction). As dispositional and institutional trust represent the generalised dimensions of trust, i.e. those not bound to specific others, they are particularly relevant in initial stages of trust establishment, i.e. in the feedforward direction, particularly early on. During initial contact with potential interactors, with whom no prior experiences have been made, deciding whether or not to interact is primarily guided those generalised constructs. Higher levels of generalised trust lead to a higher willingness to explore new alternatives for interaction over exploiting well known interactors, presenting the chance for greater innovativeness and an expansion of social interactions. However, at the same time, due to the general nature of dispositional and institutional trust, predictions about future behaviour of the trustee remain unspecific when based solely on these dimensions. By enabling interactions between yet unknown interactors, dispositional and institutional trust nonetheless represent a baseline of trust on the interpersonal level. From the basis of trust in the *general* other, trust in a *specific* other develops. The formation of a mental image of a potential interactor, and consequently the intention to trust, in turn determine the engagement of the truster in an interaction, which is a representation of trusting behavior.

The feedback direction is primarily experience driven. The outcome of each interaction represents either a positive or negative experience for the entities involved. Whether or not an interaction is perceived to

be positive or negative is dependent on the fulfilment of expectations on the part of the interactors. If a trustee has confirmed the trust put into him/her by the truster, i.e. he/she has fulfilled the expectations of the truster, a positive *interaction experience* is generated, if not a negative one.

From the generated experience, trust knowledge about the interactors is built on the interpersonal level. This does not only influence the opinion of the specific partner in a particular interaction, which in turn increases or decreases the likelihood of engaging in future interactions with that partner, but also the more general outlook with regard to institutional and dispositional trust.

The differentiation of trust, as put forward by McKnight et al., provides a structured breakdown that facilitates deriving a working definition of trust within the scope of this thesis.

2.1.2 *Focus and Trust Definition*

The different dimensions of trust presented above represent wide-ranging opportunities for study. Many of the trust models proposed in the literature (see the following Section 2.2, p. 27) consider trust as a prediction problem, specifically that of determining, or estimating, the *subjective probability of depending*. In this thesis specifically, extending estimation methods used in trust models form a core part of the scientific contribution. Therefore, the focus on trust as a prediction problem drives the working definition of trust used in this thesis. This abstracts many aspects of trust as social concept. However, it permits the use of a concise definition of trust that eliminates much of the "fuzziness" otherwise associated with the term trust.

Although sometimes deemed reductive, Gambetta [64] provides a useful definition of trust that forms the basis for a working definition for this thesis.

Definition 1 (Trust (according to Gambetta [64]))**.** *Trust* (or, symmetrically, distrust) is a particular level of the subjective probability with which an agent [*the truster*] assesses that another agent or group of agents [*the trustee*] will perform a particular action, both before he can monitor such an action (or independently of his capacity ever to be able to monitor it) and in a context in which it affects his own action.

Gambetta's definition has been adapted by a number of researchers in the field of computational trust that use it as a basis for their own trust definitions, such as Mui [151], Marsh and Dibben [136] and Jøsang [112][1]. As the focus of this thesis is on trustworthiness estimation, the particular way the term trust is used requires a closer examination and a refinement of of Gambetta's definition in order to suit our purposes.

1 For the first two, see Appendix A, pp. 227, for Jøsang's definition, see Def. 3, p. 23.

As Section 2.1.1 has illustrated, trust is a very complex concept, that has numerous facets. This complexity is mirrored in the many definitions of trust (see Appendix A) that range form definitions of trust as an action, to those defining it as a willingness to act, to those defining trust as an attitude, and to those, like Gambetta, defining it as a probability or expectation.

For probabilistic computational trust models, the core of the trust concept that is relevant in terms of mapping the mathematical to the social model, is best given by considering the interpersonal component of trust as the social counterpart to the probabilistic estimation model. In particular, a mathematical representation is given for what McKnight et al. [145] term *trusting beliefs* (see, Figure 1, p. 17). In fact, the combined representational output for trustworthiness estimates in several computation trust models, such as *CertainTrust* [173] and *Subjective Logic* [103], is called a trust *opinion*; opinions, in common parlance, are used to express beliefs.

Thus, the subjective probability of Gambetta's definition formalises (part of) the trusting beliefs of the truster with regard to the trustee's abilities. In this context, dispositional and structural trust components are *antecedents*, while trusting intentions and trust-related behaviour are *consequents*.

A sharper definition of trust can be achieved by trimming and slightly changing Gambetta's wording, in a way similar to Jøsang's definition (Definition 3, p. 23). First, the qualification that trust is *a particular level* of the subjective probability introduces unnecessary arbitrariness with regard to the interpretation of said level. Second, the focus on actions taken by the truster that affect actions of the trustee places an emphasis on activity. For the trustee's action, the term interaction, in its more abstract form discussed in Section 2.1, p. 15, can be substituted. With regard to the truster's actions that are being affected by those of the trustee, it appears to be more useful to speak of a general effect of the performance of the trustee on the truster. This stresses the fact that a (negative) effect does not necessarily as a consequence require a behavioural expression through an action.

The resulting trust definition is similar to Jøsang's (see Definition 3) [112]. It will be used as a working definition of trust in this thesis.

Definition 2 (Trust)**.** Trust is the subjective probability with which an agent [*the truster*] expects that another agent or group of agents [*the trustee*] will perform in an interaction with the truster that has an effect on the truster, in a way so that a negative outcome of the interaction will have a negative effect on the truster.

PROPERTIES OF TRUST From the definition of trust, and the structured breakdown of the term provided by McKnight et al. [142, 143, 144, 145], the view of trust as a subjective probability can be formalised further. This leads to several properties that are generally

understood to be relevant for a realisation of trust as a *computational concept* [1, 135]. The term *subjective* probability indicates that trust is local, i.e., the trust an individual A puts in another individual B is not necessarily the same that another individual C puts in B. From Definitions 1 and 2, it can also be concluded that trust is a *dyadic* and *asymmetrical relation*. That is, there are exactly two parties[2] to a trust relation, the truster and the trustee. Trust between them is directed, which means that, when considering two individuals A and B, a trust relation $A \rightarrow B$ may yield a different subjective probability, or trust value, than a trust relation $B \rightarrow A$. Additionally, trust also varies by *situation* and *context*. Furthermore, trust is *dynamic*, that is, it can change over time, and *non-monotonic*, in that the changes can be in either direction, up or down. Finally, trust is *not generally transitive*. Therefore, if A trusts B and B trusts C, it does not necessarily follow that A trusts C. Limited transitivity, however, can be achieved by employing the concept of recommendations. Recommendations provide a useful means of information sharing.

TRUSTWORTHINESS When discussing trust, the related term *trustworthiness* comes into play as well. As can be seen from Definition 2, trust has a relational quality that is directed from the truster towards the trustee. Conversely, trustworthiness is an inherent quality of the trustee. Etymologically viewed, one is tempted to believe that trustworthiness is an a posteriori measure: Trustworthiness can be seen as a measure that the trustee was *worthy of the trust* the truster put into the trustee. However, the a posteriori observation should rather be considered a behavioural expression of the inherent trustworthiness quality of the trustee. Consequently, within this thesis, trustworthiness is the, potentially unobservable, target value that is estimated by the subjective probability defined as trust in Definition 2. In other words, trust is the truster's subjective estimate of the trustee's trustworthiness, contingent on a particular situation. Therefore, a trust estimate is the outcome of a trustworthiness estimation procedure.

UNCERTAINTY Because trust is an estimate of the unobservable trustworthiness of another, ideally based on some sort of evidence, trust involves a component of *uncertainty*. From a social perspective, for example, Golembiewski and McConkie [69] have established that '*trust implies some degree of uncertainty as to outcome*'. When viewing trust as an estimation task, uncertainty in the estimate is induced by the estimation procedure, in this case bearing a close relation to the *confidence* or *credibilty* of a statistical estimate.

2 Each of these two parties may consist of one individual or a group of individuals. For the sake of simplicity, they will be considered and referred to as distinct individuals.

CONTEXT Trust is dependent on the context in which an interaction takes place. Thus, a trustee might prove trustworthy in one particular context (such as providing musical entertainment) and not in another (such as making good business decisions). The estimation and processing *mechanisms*, such as trust and certainty estimators, that are being introduced in this thesis, generally do not vary across contexts.Therefore, specific contexts are often not discussed or explicitly declared within this thesis. However, they are , implicitly always assumed; that is, any interaction between a truster A and a trustee B occurs under an arbitrary but fixed context, $\mathcal{C}$. In social settings, trust may even be transferred among contexts, say $\mathcal{C}_1$ and $\mathcal{C}_2$, as long as the truster believes these contexts to be sufficiently similar to one another.

The view of trust put forward here can be seen as somewhat reductive, as authors have argued that trust is much more than a subjective probability [36]. However, the aim of this thesis is not to provide an extensive formalisation of trust. Rather, it considers trustworthiness estimation via a trust estimate as an *estimation theoretic* problem. Therefore, epistemological discussions on the subject will be largely forgone. Nonetheless, even when considering trust a subjective probability and its establishment an estimation theoretic problem, determining said subjective probability or acting on it entails taking into account the dispositional, institutional, interpersonal and behavioural components of trust. Dispositional and institutional trust, as well as trusting beliefs are antecedents of the subjective probability of depending, and as such should be considered when modelling trust. Bayesian priors can be and are typically leveraged for encoding these in computational trust models (see Section 2.2, p. 27). The behavioural expression of trust is a consequence of a decision reached by weighing the subjective probability of depending against the willingness to depend.

To reflect the transition from subjective probability estimation to trust behaviour via a decision to depend, Jøsang et al. [112] have differentiated the trust term into two definitions, *reliability trust* and *decision trust*. While this thesis will mostly rely on the definition of trust given in Definition 2, p. 21, in Chapter 5.1, the differentiation between reliability and decision trust proves to be useful.

Definition 3 (Reliability Trust (according to Jøsang [112]))**.** Reliability trust is the subjective probability by which an individual expects that another individual performs a given action on which its welfare depends.

The definition of reliability trust covers its use as an approximator of trustworthiness. Essentially, it is a shortened version of Gambetta's definition. When having to make a decision, beyond the supposed dependability of a trustee expressed by a trustworthiness estimate, the

willingness to depend has to be considered as well. This is reflected in the definition of decision trust:

Definition 4 (Decision Trust (according to Jøsang [112]))**.** Decision trust is the extent to which a given party is willing to depend on something or somebody in a given situation with a feeling of relative security, even though negative consequences are possible.

From a modelling perspective, decision trust is generally modelled using expected utility theory [109, 131], incorporating the subjective probability and a user-defined utility function, incorporating loss and gain possible in an interaction. This also reflects the view of trusting behaviour put forward by Deutsch [44], who argues that a truster weighs the benefits of trusting another against the potential harm. Instead of loss and gain, risk and chance can be substituted here. By doing so, it becomes clear that (decision) trust can be used to quantify the expected risk or chance inherent to trusting. In comprehensive trust management systems designed to support the user in making an informed decision, it is important to convey the transition from reliability trust to decision trust to the user.

Trusters selecting a service will generally try to maximise their utility. Thus, when having to choose among several alternative providers offering equivalent services, trusters will tend to select the service with the highest expected utility EU. The expected utility function is subject to uncertainty because the true trustworthiness, p, of the trustee is, as usual, unknown.

Primarily, however, the focus of this thesis on statistical methods for trustworthiness *estimation*. In order to view trust as an estimation theoretic problem, foundational assumption with regard to the general nature of the estimation task at hand have to be established. These will be briefly motivated and introduced in the following.

2.1.3 *Assumptions*

The assumptions given in this section extend to both the contributions made in this thesis in Chapters 3, 4 and 5, as well as to the vast majority of trust models from the literature (Section 2.2. In the related work, they are generally implicitly made, without direct reference. Explicating them illustrates the common thread running through the bulk of the related work on trust models.

In this thesis, trustworthiness estimation is considered within a distributed setting. That is, in a population of individuals that interact, each individual can only rely on its own experiences with other individuals to assess their trustworthiness. While the individual can choose to share these experiences in the form of recommendations with other individuals, no global authority or distributed mechanism exists that tracks the performance of each individual to establish global trust scores.

Assumption 1 (Local Views and Incomplete Information)**.** *There exists no system-wide, global view of all interactions between all trusters and all trustees. Trust is computed subjectively by each individual truster based on local information obtained by and available to that truster.*

For the kind of trustworthiness estimation proposed in this thesis, it is important that the identity of a trustee is at least somewhat persistent. Specifically, this means that the trustee's identity exists long enough to establish a reliable trustworthiness estimate. Persistent identities are desirable for trust assessment – otherwise, a bad trust score could easily be *whitewashed* by re-entering with a new identity [54]. Also, honest and well-performing trustees would be incapable of building good trust scores if their identity were not persistent.

Assumption 2. *Trustee identities are persistent over time and do not change.*

For estimating future behaviour, a basis of knowledge that permits an informed decision has to be furnished. Extrapolation from past performance for predicting future outcomes is traditionally employed in mathematical statistics, for instance, in time-series analysis, and constitutes a mainstay assumption in trust models from the literature (see Section 2.2).

Assumption 3 (Representativity of Past Information)**.** *Information on trustworthiness in the past, for instance in the form of records of past interactions, is deemed to be indicative of trustworthiness in the (immediate) future.*

From a statistical modelling perspective, it is desirable to abstract from the continuous nature of time and assume it to be structured in discrete blocks. Interaction between individuals can be explicitly assigned to these discrete blocks of time.

Assumption 4 (Distinct, Discrete-Time Interactions)**.** *Interactions between a truster* A *and a trustee* B *are distinct events that occur at specific, discrete points in time.*

Trusters have to be able to grade interactions with trustees and record the grades assigned to the interactions, so that they can evaluate the performance of the trustees and estimate their trustworthiness over time.

Assumption 5 (Assessability of the Quality of an Interaction Experience)**.** *After each interaction between truster* A *and trustee* B*, the truster can assign a label to the interaction, which represents the quality of the interaction as perceived by truster* A*. This label is called an* interaction experience*.*

Assumption 6 (Interaction Histories)**.** *For each interaction between truster* A *and trustee* B*,* A *records* B*'s performance in a time series (or interaction history) of interaction experiences (or a sufficient statistic of that time series).*

Trusters evaluating potential trustees are primarily are supposed to be interested in predicting trustee performance in the immediate future, that is, during the next time step in the discrete-time model. Extending the goal of the prediction beyond the next time step adds uncertainty, making predictions of of trustee behaviour beyond the immediate future increasingly unreliable.

Assumption 7 (Short-Term Prediction Horizon)**.** *The goal of trustworthiness estimation is to establish trustee trustworthiness during the current time step in order to predict trustee behaviour in the immediate future, i.e., in the next time step.*

Interaction histories are modelled as discrete-time, discrete state space random processes. The discrete state space is guaranteed by limiting the parameter value that an interaction experience can take to a predefined discrete value range. Both in the related work (see Section 2.2) and in e-commerce sites, such as *eBay*[3] or *Amazon*[4], this is usual practice. *eBay* uses a binary, *Amazon* a 5-categorical model for user feedback ratings. The random distribution behind the random process is assumed to be either a binomial distribution or a m-cell multinomial distribution. This is done so that Bayesian statistics can be applied without having to estimate the distribution family from the data.

Assumption 8 (Discrete Feedback Categories for Interaction Experiences)**.** *Interaction experiences can be given as binary or m-categorical, exhaustive and mutually exclusive choices, where $m \in \mathbb{N}$ and $m \geqslant 2$.*

Assumption 9 (Random Distribution of Interaction Experiences)**.** *Each interaction experience is understood to be generated by a random process, dependent on the true but unobservable trustworthiness of the trustee involved in the interaction. Interaction experiences are assumed to be generated from a binomial or m-cell multinomial distribution.*

Assumption 10 (Expression of Trust and Trustworthiness as a Probability)**.** *Trustworthiness estimation has as its goal the estimation of one (in the binomial case) or several (in the multinomial case) p-values with $p_i \in [0;1]$. These values are represented as the parameter(s) of a binomial or multinomial distribution. The value ranges of the parameters are constrained by the properties of their respective distributions.*

The given assumptions motivate an understanding of computational trust as a binomial or multinomial probability estimation task. Appropriate statistical methods to tackle the estimation task are applied in various *computational trust models* in the literature, as introduced in Section 2.2 and are further expounded in the following chapters, especially in Chapter 3 and Chapter 4.

3 http://www.ebay.com
4 http://www.amazon.com

2.2 COMPUTATIONAL TRUST MODELS

Computational trust models are the core of various trust and reputation systems. As such, trust models provide the mechanism for the interpretation, representation and computation trust or reputation scores. As commercial reputation systems are the most common expression of trust models, it is useful to briefly differentiate trust and reputation. Following Jøsang et al. [113], reputation can be defined, in broad terms, as follows:

Definition 5 (Reputation)**.** Reputation is what is generally said or believed about a person's or thing's character or standing.

Reputation is a much more public concept than trust, which is a subjective, binary and directed relation between two individuals. Rather than being interpersonal, reputation hinges on the opinion of a multitude of individuals. As such, trust can be considered an input component of reputation, as reputation can be formed from the accumulated trust relations. But reputation is also an input component of trust, as the public opinion of peers, for instance, can influence the trust one individual has in another.

Internet users encounter reputation systems – that is, systems that accumulate and aggregate experiences or opinions from a multitude of individuals into a reputation score – in many e-commerce sites and applications. Prominent examples are *eBay*, which uses a binary feedback system, or *Amazon*, which uses a 5-categorical feedback system. These commercial systems are generally centralised, run by the respective site owners and provide global reputation scores on individuals or products. Dedicated review sites also employ centralised reputation systems, prominent examples here are, for instance, *TripAdvisor*[5] or *Yelp.com*[6], both using 5-categorical feedback, similar to *Amazon's* system. These systems generally use simple average-based point estimation techniques. Additionally, they also give the total number of aggregated experiences as additional information, in order to allow users to assess how representative the point estimate is.

In this section, the focus will be on trust models developed within the scientific community that share the assumptions outlined in Section 2.1.3. That is, the primary focus will be on related work with regard to trust models that are evidence-based, applied in a distributed setting and consider trust as a probability in $[0; 1]$. The fundamental estimation techniques underlying these models are similar to those employed in the commercial reputation systems that most Internet users are familiar with.

These Reputation systems can be considered a special instantiation of trust models. The key difference is in the origin of the data used to

5 http://www.triadvisor.com
6 http://www.yelp.com

compute a trustworthiness estimate. Trust models typically provide ways to integrate experiences a specific truster has made with recommendations received from others, that is, they combine truster endogenic with truster exogenic information. Reputation systems only rely on exogenic information, accumulating experiences from numerous contributors that report their experiences to the reputation system. In doing so, they do not apply (or require) specific mechanism for determining the quality of recommendations. Thus, they only use a subset of the capabilities provided by sophisticated trust models, such as binary and categorical estimation models.

However, the trust models presented in the following provide further methods for determining uncertainty, for enabling the propagation of trust-relevant information through recommendations and for dealing with changes in the behaviour of trustees. Other types of trust models that provide global trust scores and distributed models that compute ranks or non-probabilistic ratings will only be briefly referenced.

2.2.1 *Requirements*

The assumptions given in Section 2.1.3, in conjunction with the general research focus of this thesis on statistical methods for trustworthiness estimation, result in a number of requirements for distributed, probabilistic trust models. These requirements are motivated by a desire for accurate estimation, flexible granularity of possible feedback categories, effective use of possibly scare information, and robustness to changes in the behaviour of the involved parties.

If trust is defined as a subjective *probability* as in Definition 2, trust models that model this specific aspect of trust as a whole (compare Figure 1) require the capability to compute and represent their trustworthiness estimate from a sound statistical basis. From an estimation theoretic perspective, the computation of trust becomes a point estimation task. Its goal is the estimation of a potential trustee's trustworthiness, expressed by a time series of interaction experiences, which is assumed to follow either a binomial or multinomial distribution (Assumptions 8-10). In order to establish the reliability of the trustworthiness estimate it needs to be complemented by a corresponding estimate of the confidence in the point estimate. This follows a long-established practice in statistical estimation [157]. This second estimate models the uncertainty inherent to statistical sampling based estimation; in trust models it is commonly aggregated into a uncertainty single score that is reported alongside the trust score.

Requirement 1 (Probabilistic Computation, Representation, and Interpretation of Trust and Uncertainty from Discrete Experiences)**.** *Trust models should define trust in a probabilistic manner, so that their trustworthiness estimate* $t \in [0; 1]$*, also referred to as a* trust score*, can be interpreted*

as a subjective probability as per Definition 2. This trustworthiness estimate t *should be derived using a sound statistical estimator from a history of interaction experiences (or a sufficient statistic thereof). In order to assess the reliability of the trustworthiness estimate, a second estimate needs to be defined, modelling the inherent uncertainty of the estimation. The uncertainty estimate should be derived using a sound statistical estimation method.*

Assuming the distribution of the interaction experiences a truster records on a trustee to be distributed binomially or multinomially (Assumption 9) requires trust models processing these interaction experiences to have appropriate estimators at their disposal. In fact, trust models should provide both binomial and multinomial estimation models. By doing so, trust models are more versatile to meet the demand for different levels of granularity required with regarded to feedback categories. While some applications that use a given trust model might only require 2-categorical feedback (either a transaction went as expected, or it did not), others might necessitate a more granular categorisation to provide a grading mechanism. Since the binomial model is a specialisation of the multinomial model, providing a multinomial estimation model already incorporates providing a binomial one.

Requirement 2 (Binomial and Multinomial Estimation Model)**.** *Trust models should provide support for various degrees of feedback granularity. This capability is enabled through modeling* m*-categorical feedback with* $m \geqslant 2$*.*

Assumption 1 postulates that no view of all interactions between all trusters and trustees is necessarily accessible. For an individual truster evaluating any trustee, the truster would therefore have to rely on its own interaction experiences made in the past. In such a scenario, trust-relevant information is potentially scarce. In order to alleviate information scarcity, trust models require mechanisms that enable the sharing of trust-relevant information. These mechanisms enable one truster (acting as a *recommender*) to share its own past interaction experiences with other trusters as *recommendations* in a process called *trust propagation*. Because it is assumed that each truster has made its own experiences independently of the the interactions of other trusters, combining recommendations is simply the addition of the *endogenous information* that a truster has made itself and the *exogenous information* it has received via recommendations. Trust propagation is made more robust by providing mechanisms for assessing the reliabilty of recommenders.

Aside from recommendation-based trust propagation, information scarcity can also be tackled by incorporating sources of trust-relevant information that are not reporting independent interaction experiences. Consider, for instance, two different appraisals of the same interaction by two independent observers. The two appraisals may

differ, based on the independent views of the observers, but because the appraisals are based on observations of the same interactions, they are not independent and can, consequently, not be treated like independent recommendations. To harness non-independent trust-relevant information, trust models need to provide further mechanism for combining trust sources.

Requirement 3 (Mechanisms for Trust Propagation and Combining Trust Sources)**.** *Trust models should provide mechanisms for sharing trust-relevant information obtained by several trusters independently of each other through recommendations, i.e, methods for* trust propagation*. Recommendations, representing* exogenous information *not collected by the truster itself, have to be integrated with* endogenous information *of the truster. Appropriate methods for assessing the reliability of recommendations should be provided. Further methods for combining trust-relevant information from non-independent sources should also be provided.*

Definition 2 considers trust a *subjective* probability. One component of this subjectivity is given by the fact that each truster has its own, individual interaction experiences with a trustee and thus its evaluation of said trustee might differ from that of other trusters. Another component of subjectivity, however, is injected into trustworthiness estimation by the other aspects of trust (see Figure 1), aspects that do not reduce trust to a frequentist point estimate. Components such as ones disposition to trusting in general, trust in structural assurances, as well as individual belief systems, can affect the estimate of trustworthiness. The estimation model used for assessing trustworthiness should therefore be able to account for non-frequentist and non-experience-based factors in the derivation of a trust score. Particularly during system bootstrapping, when no or little interactions have taken place and interaction histories are only sparsely populated with information, non-experience-based factors of trust are necessary to kickstart trust building.

Requirement 4 (Mechanisms for Integrating Non-Frequentist Information)**.** *Trust models should provide mechanisms for including trust-relevant information that is not directly derived from interaction experiences. By doing so, dispositional/generalised and institutional/structural components of trust, as well as trusting beliefs (see, Figure 1) are integrated into the trust estimation.*

For statistical reasons, trustee behaviour is modelled by assuming binomial or multinomial distributions of interaction experiences. This abstraction is useful, as it justifies the usage of point estimation techniques, and, coupled with the assumption of *stationary behaviour* over time, guarantees the convergence of the estimator to the true, unobservable value of a trustee's trustworthiness. Unfortunately, stationarity may be too strong an assumption in the real-world, as behaviour

tends to change over time. Therefore, trust models should compensate for the possibility of changing, dynamic trustee behaviour.

Requirement 5 (Mechanisms for Coping with Changing Behaviour)**.** *Trust models should provide methods for compensating the effect on estimation accuracy of changes of trustee behaviour over time.*

In the following Section 2.2.2, distributed, probabilistic trust models will be presented that are most closely related to the research presented in this thesis. Their compliance to the requirements listed in this section will be considered and possible room for improvement will be briefly stated.

2.2.2 *Distributed, Probabilistic, Evidence-based Trust Models*

The focus in this section will be on models that estimate the trustworthiness of a potential trustee based on a truster's direct evidence from past experiences or recommendations. The application fields for which the presented models were developed or are applicable for wide-ranging, from e-commerce [41], agent systems [94, 198], to online social networks [68, 151], to ubiquitous computing [121, 172], to ad-hoc and sensor networks [185] and P2P systems [29, 43], for instance.

The actual estimation model underlying a trust model should be agnostic about the application field it is used for. Thus, while there are some application specific trust models among those introduced in the following, they will be considered for their trustworthiness estimators without particular attention being paid to the application motivating their design. *CertainTrust* [173] and *CertainLogic* [75, 175] form the basis for the extensions proposed in this thesis. They will be discussed in Section 2.2.2.1. *CertainTrust/CertainLogic* is derived from *Subjective Logic* [103], a comprehensive framework for reasoning about trust under uncertainty, discussed in Section 2.2.2.2. Other prominent probabilistic trust models will be briefly discussed in Sections 2.2.2.3 to 2.2.2.10

2.2.2.1 *Ries'* CertainTrust/CertainLogic *model*

CertainTrust is a trust model proposed by Ries [173], derived from Jøsang's *Subjective Logic*, to which it is isomporphic. *CertainTrust* represents trust and its inherent uncertainty in the form of *CertainTrust opinions*, parameter triples $\omega := (t, c, f)$, with

- $t \in [0; 1]$, a trustworthiness estimate based on the *Maximum Likelihood Estimator* $t = \frac{r}{r+s}$, where r is the sum of all positive experiences with the trustee under evaluation and s the sum of all negative experiences with the trustee,

- $c \in [0;1]$, a certainty estimate representing the confidence in t, and
- $f \in [0;1]$, an initial trust value that can be used to incorporate dispositional or other non-frequentist components of trust into the model.

Feedback in *CertainTrust*, that is, the categories in which experiences are recorded, is binary, resulting in a binomial estimation model. The maximum likelihood estimation technique, using the arithmetic mean, is a well-defined and conventionally accepted method for point estimation in binomial probability estimation tasks. In case of no evidence, i.e., $r = s = 0$, t is defined as 0.5.

CertainTrust's certainty estimate c models uncertainty in the interval $[0;1]$, dependent solely on the number of experiences. When no evidence is available, $r = s = 0$, uncertainty is maximised, and the certainty parameter $c = 0$; once a pre-determined, representative number of experiences $N \in \{\mathbb{N}, \infty\}$, uncertainty is minimised and $c = 1$. The certainty measure is computed in a somewhat ad-hoc manner as:

$$c = \frac{N \cdot (r+s)}{2 \cdot w \cdot (N - (r+s)) + N \cdot (r+s)}$$

The parameter $w \in \mathbb{R}^+$ is a system parameter with a default value of $w = 2$. Combined with the initial trust value f, it is used for mapping the initial trust value of *CertainTrust* to a *Bayesian prior*[7].

From the representation trust as $\omega := (t, c, f)$, a single score, aggregate expectation value can be computed as $E(t, c, f) := c \cdot t + (1 - c) \cdot f$. This provides a compact representation of the estimation output, that shifts with increasing certainty from an initial, dispositional component of trust, f, to the evidence-based, probabilistic estimate t

The representation of trust in *CertainTrust* is amenable to a estimation theoretic interpretation. With regard to the trustworthiness estimate t, the demands of Requirement 1 are adequately met. In principle, the certainty estimate c can be interpreted in the same manner. However, the actual estimator used is not based on a conventional statistical measure of confidence in an estimate. By replacing the certainty estimator with another estimator with a stronger estimation theoretic foundation, the *CertainTrust* representational model can be made compliant with Requirement 1, readily.

With regard to Requirement 2, calling for both binomial and multinomial estimation models to be supported by trust models, *CertainTrust* provides only binomial estimation. Both the representation, as well as the computation of trust have the be extended.

In its basic form *CertainTrust* provides robust trust propagation mechanisms for its binomial trust representation. These account for

7 For more details on the mapping of f and w to Bayesian priors, see Chapter 3.1.6, p. 66.

potentially dishonest recommenders and are designed to counter sybil attacks. Extended with *CertainLogic* [75, 175], *CertainLogic* gains capabilities to combine trust-relevant information from non-independent sources, through the introduction of various averaging fusion operators. Requirement 3 is thus fulfilled for binomial *CertainTrust* opinions. In order to enable multinomial capabilities for combining trust-relevant information, *CertainTrust* has to be suitably extended. Alternatively, once the representational aspect of the *CertainTrust* model has been extended to handle multinomial estimation, the isomorphism to *Subjective Logic* can be leveraged to access its capacities for combining information.

By integrating parameters $\mathfrak{f}$ and w, *CertainTrust* can integrate non-frequentist information in its estimation. Additionally, *CertainLogic* operations can be used to fuse frequentist and non-frequentist information if a mapping from non-frequentist information to *CertainTrust* opinions can be provided. For the binomial case, *CertainTrust* fulfils Requirement 4, while extensions to the model have to guarantee its adherence in the multinomial case.

CertainTrust also provides an approach for dealing with changing trustee behaviour over time, meeting Requirement 5. It does so by introducing a robust ageing operation that fades out older information in favour of more recent experiences. The ageing employed in *CertainTrust* is improved compared to the basic version employed in other trust models, resulting in a lower estimation error when using this particular ageing operation. The ageing approach, even in this modified version, still discards information that might still be relevant and useful for trustworthiness information.

Due to its representation of trust and (un-)certainty estimation that is highly amenable to an estimation theoretic interpretation, *CertainTrust* represents a good foundation for incorporating extensions rooted in statistics and estimation theory. *CertainTrust* also provides a connection to Bayesian statistics, which can be leveraged in several ways; this connection is not only particularly useful for designing a *Multinomial CertainTrust* model, but can motivate and justify refinements to the original binomial *CertainTrust* model, for instance, in the design of statistically well-founded certainty estimators.

The contributions presented in Chapters 3, 4, and 5 are modelled as extensions to the *CertainTrust* model, although they are, with only slight modifications, generally applicable to other trust models.

2.2.2.2 *Jøsang's* Subjective Logic *and Corresponding Reputation Systems*

Subjective Logic, introduced by Jøsang in [103], is a prominent framework for reasoning over uncertain probabilities. It serves as the basis for both the *Beta Reputation System* [108], a trust model for binomial trustworthiness estimation, and the *Dirichlet Reputation System* [107], the multinomial generalisation of the *Beta Reputation System*. *Subjec-*

tive Logic also served as the direct inspiration behind the development of *CertainTrust* and *CertainLogic*, as can be evidenced by the maintenance of isomorphisms between *CertainTrust*/*CertainLogic* and corresponding operations in *Subjective Logic*.

While *CertainTrust*'s representation of trust is more geared towards an estimation theoretic approach, *Subjective Logic* is rooted in Dempster-Shafer belief theory [183]. For this, Jøsang provides a mapping from the so-called *evidence space*, in which past experiences are recorded as sums of all positive experiences, r, and negative experiences, s, to a so-called *opinion space*. This mapping, essentially using mechanisms of Bayesian probability theory, maps r and s to a tuple (b, d, u), where $b \in [0;1]$ the belief in a positive outcome, $d \in [0;1]$ the belief in a complementary negative outcome, and $u \in [0;1]$ the inherent uncertainty. The value ranges of the three parameters are further constrained by having to conform to $b + d + u = 1$. Additionally, a parameter $a \in [0;1]$ models a *base rate* that can be used to incorporate dispositional or other non-frequentist components of trust

For the binomial case, the mapping is as follows:

- $b \in [0;1]$, essentially a trustworthiness estimate based on a modified *Maximum Likelihood Estimator* $b = \frac{r}{r+s+W}$, where r is the sum of all positive experiences with the trustee under evaluation and s the sum of all negative experiences with the trustee, and W a system parameter defaulting to $W = 2$,
- $d \in [0;1]$, essentially an *un*trustworthiness estimate based on a modified *Maximum Likelihood Estimator* $b = \frac{s}{r+s+W}$, where r is the sum of all positive experiences with the trustee under evaluation and s the sum of all negative experiences with the trustee, and W a system parameter defaulting to $W = 2$,
- $u \in [0;1]$, an uncertainty estimate representing the confidence in b and d, computed as $u = \frac{2}{r+s+W}$.

Obviously, there is a dependence among the parameters, as the constraint of $b + d + u = 1$ means that if one parameter changes, the other two parameters have to change as well in order to fulfil the constraint. This can have unintuitive results for those acquainted more closely with a probabilistic representation than with a belief representation. [78] illustrates this well:

- assume one has collected one positive and one negative experience on a trustee; this yield $(b = 0.25, d = 0.25, u = 0.5)$. The belief would therefore be $b = 0.25$ that the trustee is trustworthy. Contrast this with the frequentist probability of 0.5 that a Maximum Likelihood Estimator would have computed, albeit with a very high uncertainty;
- now assume that the number of collected experiences has increased to ten positive and ten negative experiences; this yields

$(b = 0.\overline{45}, d = 0.\overline{45}, u = 0.0\overline{9})$. Here, with increasing evidence, the belief approaches the maximum likelihood estimate, which would still be 0.5 but with a lower uncertainty.

As can be seen from the example, the belief-based representation of *Subjective Logic* is semantically different from a purely probabilistic interpretation. Users of *Subjective Logic* therefore have to be made familiar with belief theory, in order to be able to use and appreciate the way that information is conveyed here. Additionally, the uncertainty function appears to be defined in a rather ad-hoc manner. Nonetheless, it has a direct impact on the value of the belief value b, complicating the interpretation of the belief-value further. When considering trust a *subjective probability*, as it is within this thesis (Definition 2), rather than a belief, a more estimation-oriented representation is preferred. The estimation model, particularly with regard to uncertainty estimation, only meets Requirement 1 to a limited degree.

Over the years, *Subjective Logic* has been extended considerably, for instance, in [114] to handle multinomial feedback (thus meeting Requirement 2), and has been tweaked in its parameters. [106] provides a summary of the current state of the ongoing development of *Subjective Logic*. This includes various ways to aggregate independent and non-independent trust sources, so that *Subjective Logic* fulfils Requirement 3. With regard to Requirement 4, the base rate parameter a permits the integration of non-frequentist information into the estimation model. As is the case for *CertainTrust*, changes in trustee behaviour are compensated by applying ageing. While ageing has the shortcoming of discarding information indiscriminately only based on its age, Requirement 5 is largely addressed.

Overall, *Subjective Logic* provides a comprehensive framework for reasoning based on belief theory. It also provides most of the basic mechanisms required by a trust system. Room for improvement exists with regard to its uncertainty estimator, an possibly with regard to some of its mechanisms for estimating recommender trustworthiness and for coping with changing trustee behaviour. Also, its belief logic inspired representation does not lend itself to the statistical interpretation of trust desired in this thesis as the one provided in *CertainTrust*.

In the following, several other prominent trust models will be briefly introduced.

2.2.2.3 *Buchegger's model*

Buchegger et al. [28, 29] proposed a robust trust model for P2P environments and wireless ad-hoc networks. Its focus is on identifying false recommendations, which is achieved by a so-called "deviation test" that determines the similarity between a recommendation and the truster's own experience. The test compares the expectation

values of the recommendation and the truster's experience; the test succeeds if the difference of the expectation values is below a predetermined threshold. Unless the test succeeds, the recommendation is not included in the trust computation.

Buchegger's estimation model is based on a modified Bayesian estimator, essentially a posterior Maximum Likelihood approach, used for the estimation of a binomial trustworthiness estimate. Uncertainty estimation is not explicitly included in a statistically sound way and multinomial support is not given. The model provides methods for including recommendations from independent sources, but does not address aggregation of non-independent information. The scope of the model is largely on frequentist information and does not address the integration of non-frequentist information explicitly, although it can still be integrated via a Bayesian prior. Basic ageing is implemented to cope with changing trustee behaviour, but affects the estimation quality adversely (as shown in [173]).

2.2.2.4 *Despotovic and Aberer's model*

Despotic and Aberer propose a simple Maximum Likelihood-based trustworthiness estimation model in [42] and investigate probabilistic and social network trust in [43]. They do not provide a comprehensive trust model per se; rather, they investigate the predictive accuracy of Maximum Likelihood and Bayesian estimation in P2P environments, also including recommender collusion. Aside from standard binomial trust models, they also investigate the estimation of normally distributed feedback. Uncertainty is not modelled as statistic parameter. The need for a multinomial estimation model and the demands of Requirements 3 to 5 are not addressed.

2.2.2.5 *Huynh's* FIRE *model*

Huynh designed the *FIRE* model [93, 94] for application in multi-agent societies. *FIRE* combines a wide variety of information sources in a modular approach. The model includes four distinct types of modules contributing to the computation of a trust score. The four modules presented in [93] are as follows:

- *Interaction Trust*: computes trust from the endogenous information the truster has on a trustee, i.e., its own prior experiences.
- *Role-based Trust*: computes trust from the role that the given trustee fulfils within the interaction at hand.
- *Witness Reputation*: computes a trust score from the exogenous recommendations from others.
- *Certified Reputation*: computes a trust score from certificates supplied by the trustee to the truster, in which other trusters certify the trustee's trustworthiness.

FIRE assumes trustee behaviour to be distributed according to a continuous, Gaussian distribution in $[-1;1]$. Consequently, its estimate is also in $[-1;1]$, and therefore not a probability, but rather exemplifies the expected quality a truster can expect from a trustee in an interaction, scaled to $[-1;1]$. As such, *FIRE* is not a *probabilistic* trust model. However, its extensible nature is of some interest. *FIRE* combines its different modules by computing a trust score and a reliability score for each component. Aggregate scores are subsequently computed for trust and reliability. The trust score is computed as a weighted sum of the component trust scores, weighted by their reliability and a predetermined, component-specific weight. The estimation procedure is sound in general, given an assumption of Gaussianity (which is difficult to enforce in real-world settings). However, *FIRE* trust scores are not probabilistic, therefore fail to meet Requirement 1. Similarly, *FIRE* does not address categorical feedback, and therefore does not address Requirement 2. *FIRE* does support trust propagation and, in a manner, the combination of trust sources through its modular approach, meeting Requirement 3. It is also flexible enough to include non-frequentist information (Requirement 4); its certified reputation component, for instance, can be considered non-frequentist information. Further such components are conceivable. In order to deal with changing trustee behaviour, the trust model implements ageing.

FIRE integrates and adapts aspects of other trust models for agent-based environement, such as Sabater's *REGRET* model [179] and Ramchurn's model [168]. Due to its assumptions of Gaussian feedback, it is, however, not a probabilistic model in the sense that it interprets trust as a *subjective probability*. Furthermore, its reliability estimate is functional but ad-hoc, as is the integration of its components.

2.2.2.6 *Kinateder's* UniTEC *model*

With *UniTEC* [119, 120, 118], Kinateder introduced a framework that outlines a number of generic requirements for trust models.In this framework, identifies five main aspects of trust estimation:

- *Trust Measure*: The trust score assigned to a trustee, which should model the trustees trustworthiness from complete distrust, to neutral (i.e., mistrust), to complete trust.
- *Trust Certainty*: A confidence measure on the trust score.
- *Trust Context*: The general context in which a truster trusts a trustee.
- *Trust Directness*: Kinateder distinguishes two categories here: *direct* and *indirect* trust. The former refers to direct interactions between truster and trustee, while the latter concern referrals and recommendations.

- *Trust Dynamics*: Trust changes with the addition of new evidence or as time passes.

The strength of *UniTEC* lies in providing an explicit framework for what a trust model should be able to compute. *UniTEC* provides mappings to integrate trust models such as [1, 179] into its framework, with additional constraints where necessary. Except for *trust context*, which is implicitly assumed within this thesis but not explicitly modelled, the aspects of Kinateder's model have influenced the assumptions and requirements listed in Sections 2.1.3 and 2.2.1.

In its original publication [119], *UniTEC* uses a simple update rule for trustworthiness estimation, derived from geometric learning. This rule already incorporates ageing and fades information over time. Despite its derivation, it is interpreted as a probability. For a certainty measure, the use of the reliability measure from *REGRET* [179] is proposed. While both trust and certainty are interpreted in a probabilistic manner, the manner of their computation is somewhat ad-hoc. Thus, while functionally useful, the goal of their design is not statistical accuracy. Requirement 1 is therefore not fully met by *UniTEC*.

Similar to *REGRET* [179] and *FIRE* [94], the *UniTEC* model deals with continuous rating, instead of m-categorical feedback. While this is easy to model using a Gaussian distribution, Gaussianity cannot be guaranteed in the real-world, theoretically requiring the estimation of the distribution of the actual distribution underlying the continuous feedback. Consequently, Requirement 2 is not met.

UniTEC provides a number of mechanisms for trust propagation, as demanded in Requirement 3. Kinateder investigates logic operators for finding sources of trust relevant information by chaining recommendations. Averaging operations for aggregating dependent information are not presented in *UniTEC*, however. Requirement 4 is also not addressed, as the model is geared towards derivation of trust from evidence, without explicitly accounting for non-frequentist information.

UniTEC uses fading and aging mechanisms to cope with with changing trustee behaviour. Requirement 5 is thus generally fulfilled.

2.2.2.7 *Mui's model*

Mui proposed an early Bayesian model [152, 151] for the decentralised computation of trust. It is built on a sound statistical foundation, based on posterior Maximum Likelihood estimation over a Beta distribution. The certainty measure is computed as a function of a predefined desired number of experiences and the actually collected number of experiences, based on the computation of a *Chernoff Bound* on the number of experiences necessary to achieve a desired level of confidence.The estimation model computes and interprets trust as probability, the certainty parameter is derived in a statistically mean-

ingful way (Requirement 1). The estimation model, however, is defined only for binary feedback, not considering m-categorical feedback for $m > 2$. Multinomial support, therefore, is not given.

Um considers recommendations and trust propagation within his model, he does not, however, provide explicit mechanisms for integrating endogenous and exogenous trust-relevant information, either independent or non-independent. The model, therefore, does not meet Requirement 3. However, by embracing Bayesian estimation, Mui permits for the integration of non-frequentist information as part of the Beta prior (Requirement 4). Changes in trustee behaviour are not explicitly considered in Mui's model.

2.2.2.8 *Teacy's* TRAVOS *model*

TRAVOS [189] is a Bayesian trust model for binary feedback. *TRAVOS* leverages posterior Maximum Likelihood estimation for computing a trustworthiness estimate, as well as a statistically well-defined certainty estimate based on a Beta posterior probability density function. The computation, representation and interpretation of trust as a probability is given for this model (Requirement 1). It is, however, geared exclusively towards binary feedback, not accounting for multinomial feedback.

TRAVOS places great importance on reliably incorporating recommendations into its estimation model. For this, it provides a sophisticated mechanism to assess recommender reliability. It does not provide mechanisms for aggregating non-independent information, though. Thus, it fulfils Requirement 3 only partially. Just as Mui's model, however, the presence of a Bayesian prior enables the integration of non-frequentist information into the estimation process, in principle meeting the demands of Requirement 4. Changes in trustee behaviour are not considered.

2.2.2.9 *Teacy's* HABIT *model*

HABIT [190] provides a hierarchical model for estimating and combining trust from different sources. *HABIT* uses a hierarchical Bayesian model that is based on sound statistical principles to combine various forms of information in order to compute a trust score. A very interesting feature of using a hierarchical Bayesian approach is the ability of *HABIT* to combine reputation and experience from two different models, for instance a categorical model for experience and a Gaussian model for reputation. *HABIT* also relies on expected utility theory in its formulation of a trust decision.

The focus of *HABIT* lies on sound statistical principles for the combination of trust information from different sources. While this bears some similarity to the work presented in this thesis, the goal and ap-

proach varies. For instance, *HABIT* does not seek to model trust as a subjective probability per se, in the way presented in this thesis.

2.2.2.10 *Wang and Singh's model*

Wang and Singh propose a probabilistic, evidence-based model [196, 197, 198] for trustworthiness estimation. Their approach focuses on Bayesian estimation of trustworthiness from binary feedback, harnessing a Bayesian posterior Maximum Likelihood Estimator to compute a trust score. They also provide a probabilistic certainty estimator, based on a Beta posterior probability density function. In their design of this certainty estimator, they postulate a number of useful properties for certainty estimation [196] that will be used in this thesis (Chapter 3). The estimation model meets Requirement 1, it does not, however, extend to multinomial estimation, thus not meeting Requirement 2.

Wang and Singh's model proposes a number of different sophisticated methods to assess recommender trustworthiness [198]. These are tested against similar mechanisms from the literature, such as those provided in *TRAVOS* or the *Beta Reputation System*. The evaluation, however, uses an unconventional scenario to evaluate the effectiveness of the approaches; recommender trustworthiness is evaluated not after every interaction, but after a batch of interactions, thereby increasing the available information during recommender evaluation considerably. Chapter 4 evaluates Wang and Singh's measures for recommender trustworthiness, among others, against a novel method introduced in this thesis, in a more conventional setting, in which recommender trustworthiness is evaluated more frequently. Since the model does not provide any methods for aggregating non-independent information, Requirement 3 is only partially met.

By using Bayesian estimation, non-frequentist information can be integrated into the prior of Wang and Singh's model. While this is not explicitly modelled, Requirement 4 is implicitly fulfilled. Their model does not take changing trustee behaviour into account (Requirement 5).

2.2.2.11 *Further Experience-Based Trust Models*

The trust models introduced in the previous Sections 2.2.2.1 to 2.2.2.10 are prominent examples of numerous other probabilistic trust models. Many more models exist, most, however, adapt one or more of those introduced here to specific applications. Examples of other models include, among many others, Ganeriwal's model [65] for wireless sensor networks, Billhardt's [18] and Hang's [79] models for service selection, as well as the prominent trust model *EigenTrust* [115] by Kamvar et al., which computes a global trust score for each participant in

	Requirements				
	1	2	3	4	5
CertainTrust	partial	binomial	+	+	basic
Subjective Logic	partial	+	+	+	basic
Buchegger	partial	binomial	partial	+	basic
Despotovic & Aberer	partial	binomial	-	-	-
FIRE	-	-	+	+	basic
UniTEC	-	binomial	partial	-	-
Mui	partial	binomial	-	+	-
TRAVOS	+	binomial	partial	+	-
HABIT	partial	+	partial	+	-
Wang & Singh	+	binomial	partial	+	-

Requirement 1: Probabilistic Computation, Representation and Interpretation of Trust and Certainty
Requirement 2: Binomial and Multinomial Estimation Model
Requirement 3: Trust Propagation and Combining Trust Sources
Requirement 4: Integration of Non-Frequentist Information
Requirement 5: Changing Trustee Behaviour

Table 1: Trust Models – Fulfilment of Requirements 1-5.

a P2P environment. A number of surveys exist that provide useful overviews over the numerous trust models [8, 76, 113, 195, 207].

2.2.2.12 *Fulfilment of Requirements*

Table 1 summarises the extend to which the trust models described in Section 2.2.2 meet the requirements postulated in Section 2.2.1. The entry for *CertainTrust* [173] encompasses the extended version of *CertainTrust*, i.e., its combination with *CertainLogic* [175, 75]. In a similar manner, the entry for *Subjective Logic* [103] encompasses the *Beta* [108] and *Dirichlet Reputation Systems* [107]. As can be seen, none of the trust models meet all requirements:

- Requirement 1 is met partially by most models in the sense that they provide appropriate probabilistic trustworthiness estimators but do not provide adequate estimators for certainty estimation; only *TRAVOS* [189] and Wang and Singh's model [197, 194] offer statistically well-derived certainty estimators.
- Requirement 2 is also only partially met by most models; except for *FIRE* [94], which uses a Gaussian estimation model, all the trust models provide binomial estimation models, how-

ever, only *Subjective Logic* provides a multinomial generalisation. *HABIT* can integrate any kind of estimation model.

- Requirement 3 is fully addressed by both *CertainTrust*/*CertainLogic* and *Subjective Logic* through a wealth of operators for combining both independent and non-independent trust sources. *FIRE* is, by its modular nature, at least in principle capable of doing the same. Partial fulfilment of Requirement 3 means that the trust models have facilities for trust propagation, i.e., for combining independent trust sources.

- Requirement 4 is in principle met by all models that harness Bayesian estimation in their estimation model; here, the prior information can encode subjective and non-frequentist information. The *FIRE* model is ablate achieve the same feat by providing additional modules that model subjective or non-frequentist information.

- Requirement 5 is generally met, by those models that address it, through the application of ageing or fading mechanisms. These enable the models to cope with changing trustee behaviour but limit the accuracy that can be achieved by their trustworthiness estimators. Therefore, the capabilities for dealing with changes in behaviour are denoted as *basic* in Table 1.

In Chapters 3 and 4, methods are developed that aim at enabling trust models to simultaneously meet all five requirements. For this, the *CertainTrust* model will serve as a basis.

2.2.3 *Steoreotyping Trust Models*

A more recent research trend in trust models has focussed on the application of feature-based supervised prediction methods to trustworthiness estimation. By introducing a set of features that a potential trustee exhibits, supervised machine learners can be trained to find correlations between the exhibited features and trustee trustworthiness, potentially mitigating the need to collect a representative number of experiences for *each* trustee in order to derive individual trustworthiness estimates. By using feature-based estimation, trust may thus be bootstrapped from a smaller number of interactions than would previously have been necessary. Stereotyping trust models are themselves derived from stereotype-based user modelling, pioneered by Rich [171].

Research on trust bootstrapping via stereotyping by Liu et al. [128], Burnett et al. [31, 32], and Fang et al. [52] is directed at providing a better generalisation ability by creating stereotypical profiles for generalisation in agent societies. Through their use of machine learning

and data mining, this research can be considered closely related to the work presented in Chapter 5.2.

LIU'S *stereotrust* MODEL The approach by Liu et al. [128] leverages *linear discriminant analysis* as a classifier for the prediction of an outcome. The outcome of this prediction is a class label, indicating whether or not the next interaction between an evaluating truster and a potential trustee is expected to be positive or negative, a concordant confidence score is given as the probability of the expected outcome. This differs from the view of trust as a probability. In fact, the confidence score represents the estimated trust score in probabilistic methods. Thus, the model by Liu et al. does not account for estimation uncertainty in determining the probability of a positive outcome. Additionally, *linear discriminant analysis* is a rather weak estimator, requiring a strong correlation between feature set and trustworthiness score to be effective. In [128], this is guaranteed by the choice of data sets used for the evaluation but can, in all likelihood, not be guaranteed in real-world settings.

BURNETT'S STEREOTYPICAL TRUST MODEL Burnett et al. [31] use *M5 decision trees* [26, 166] as their base learners. This approach is integrated with Jøsang's *Subjective Logic* and uses *Subjective Logic* opinions as its regressands,. The model-building regression trees are capable of returning a probabilistic trust score in the setting chosen by Burnett et al., which involves binary feedback, concordant with a binomial trust model. The trust score, albeit without a certainty measure, is then used as a *base rate* – that is, dispositional trust information – within the framework of *Subjective Logic*. In simulations, Burnett et al. show that this leads to a markedly improved trustworthiness estimation quality. However, they state that this is very much dependent on the discriminative power of the provided features. While the M5 tree is considerably more capable than the linear discriminant analysis classifier in in Liu et al.'s model, the simulations in [31] thus only show the principle practicability of the approach. An investigation of how supervised machine learners perform with real-world trust and reputation data is still missing.

FANG'S GENERALISED STEREOTYPICAL TRUST MODEL Fang et al. [52] use *Fuzzy Semantic Decision Trees* as base learners and derive a probabilistic trust score based upon nominal features – handled by a so-called semantic process – and non-nominal features – handled by a fuzzy process. The fuzziness is introduced as it is supposed by the authors that fuzzy learning techniques perform better under limited data. Furthermore, an ontological knowledge representation is chosen for the nominal features, which are placed in a structure that allows for generalisations. This provides the learner with additional

information, albeit at the cost of ontology creation. The system is evaluated through simulation an compared to the approach by Burnett et al. Again, not evaluation against real-world data has been conducted. These simulations demonstrate the general feasibility of the approach. They do not, however, show the capability of the base learner and the entire approach to build a prediction model from data that is has *not* been carefully generated to show efficacy of the approach in a simulation settings.

In Chapter 5.2, a number of supervised machine learners will be applied to a real-world data set generated by a reputation system. Contrary to the simulated data in [31, 52], the correlation structures between regressands (i.e., trust scores) and regressors (i.e., features) are not pre-determined or even certain to even exist. The results should be indicative of whether or not stereotyping methods and reputation-based selection are readily compatible.

2.3 FURTHER STATISTICAL METHODS IN TRUSTWORTHINESS ESTIMATION

In this thesis, a number of statistical methods are applied, ranging from estimation theoretic concepts in the field of point estimation, to exact hypothesis tests, to non-parametric, model-free learning machines. In Chapter 3, Bayesian point and interval estimation techniques are put to use for determining trustworthiness and certainty estimates. Here, standard statistics textbooks furnish background on point estimates via binomial and multinomial proportions, specifically, Agresti on categorical data analysis [3], Berger on statistical decision theory and Bayesian analysis [12], Bernardo and Smith on Bayesian theory [13], Bolstad on Bayesian statistics [20], Casella and Berger on statistical inference [35], Jeffreys on the theory of probability [99], and Lehman and Casella on the theory of point estimation [126]. Further work in Chapter 3 on interval-based certainty estimation is built upon the work on interval estimation of binomial proportions by Brown et al. [27], supported by seminal work by Jaynes [96, 97, 98]. In order to extend interval-based certainty estimation to multinomial trust models, simultaneous confidence intervals were adapted for use in multinomial certainty estimators, relying on prior work by Goodman [70].

Chapter 4 introduces multinomial variants of various operators in *CertainTrust* and *CertainLogic*, extending work by Ries [173, 175], as well as our own work in [77]. The original operators, in turn, are based on work by Jøsang [103]. Also in Chapter 4, hypothesis testing using exact hypothesis tests are adapted in several applications, namely, the determination of recommender trustworthiness, the computation of the degree of conflict in opinion fusion and change point detection for changing trustee behaviour. *Fisher's Exact Test* [56, 55]

for binomial models and the *Fisher-Freeman-Halton Test* [58] for multinomial models are used to derive similarity scores for the comparison of opinions. These tests also represent the basis for the change point detection model [163] that is applied to trustworthiness estimation, specifically from work by Ross et al. [178].

In Chapter 5, non-parametric, model-free supervised learning machines are applied to trustworthiness estimation. The work by Malley et al. [133] on consistent supervised estimation served as a major inspiration in this chapter, in particular with regard to the supervised learning machines that were chosen. Specifically, these included CART [26] and M5 [166] decision trees, random forests [25], as well as a k-nearest-neighbour variant [24]. Most of these are from Breiman's extensive and seminal works on supervised estimators [26, 24, 25]. Further references are given – where necessary – throughout this thesis.

2.4 CHAPTER SUMMARY

In this chapter, background information and related work on trust and trust models have been presented. Three subdivisions have been made, first presenting trust and trustworthiness-related concepts, second going into computational trust models, before closing with a brief overview of the most relevant statistics work used in this thesis.

- *Concepts of Trust and Trustworthiness*: Through differentiating trust into its various aspects [143, 144] in Section 2.1, the view of computational trust as a subjective probability is motivated and a workable definition of trust, adapted from Gambetta's popular definition [64], is introduced. From this, the need for estimation theoretic tools for determining such a subjective probability is derived. In order to define appropriate estimators in later chapters, assumptions with regard to the goal and nature of the trustworthiness estimation task at hand are given an explication in Section 2.1.3.

  By taking these steps, the way that trustworthiness estimation fits into the complex concept of trust is established. In computational trust, trustworthiness estimation functions as an analogy to what McKnight et al. [143] term *trusting beliefs* in social trust. This defines the type or aspect of trust considered in this thesis and provides a base for the formalisations given in the coming chapters.

- *Computational Trust Models*: Section 2.2 starts with a postulation of requirements to be met by state-of-the art trust models in Section 2.2.1. In Section 2.2.2 those trust models that are most closely related to the work presented in this thesis are intro-

duced and their compliance with the requirements outlined in Section 2.2.1 is discussed. None of the trust models in Section 2.2.2 meet all the postulated requirements. While no one-fits-all solution currently exists, a number of trust models provide a solid basis for further extensions, specifically *Subjective Logic* and *CertainTrust*/*CertainLogic*. Other trust models provide interesting solutions to specific requirements, for instance, *TRAVOS* and Wang and Singh's model, which integrate sophisticated certainty estimators. Stereotyping trust models are introduced, as they are relevant in Chapter 5 for their application of supervised learning to trustworthiness estimation.

By explicitly giving the assumptions that the mathematical formalisations in this thesis are building upon, the approach for modelling trust and certainty estimators, for applying processing steps and developing further extensions is given a clear basis and delineation. Deriving the requirements that drive the development of trustworthiness estimation techniques and extensions in this thesis, clear goals for what a trust model should be capable of are formulated, both motivating and allowing for a directed development approach. A comparison of various existing trust models regarding the requirements reveals that individual models (are designed) to meet individual requirements, but none meets all requirements simultaneously.

- *Further Statistical Methods*: A wider view of related work is presented in the final Section 2.3 of this chapter. It lists the literature from statistics and machine learning that provides the basic tools to be applied throughout the remainder of this thesis.

  By providing a brief overview over the methodological basis in statistics and machine learning, the roots of trustworthiness estimation in those fields are illustrated. This is done in order to convey that trust models (should) build upon established and seminal work.

In Chapters 3 and 4 mechanisms will be introduced and adapted to *CertainTrust* that will allow this model to meet all the requirements of Section 2.2.1. This includes statistically well-founded trustworthiness and certainty estimators for both binomial and multinomial estimation models. Furthermore, novel methods for integrating and combining trust-relevant information, and for dealing with changing trustee behaviour will be developed. All of this is done with the help of statistical tools (compare, Section 2.3).

In Chapter 5, methods for generalising trustworthiness information will be investigated. Of particular interest for this is the application of supervised machine learners. Over the past years, several stereotyping trust models have been proposed, as introduced in Section 2.2.3. In the literature, these stereotyping models are evaluated by

simulations that assume a reasonably benign distribution of the data that is used for training. In Chapter 5.2, supervised methods will be tested against a real-world data set, in order to find out whether the real-world distribution of the data generated by reputation systems is amenable to supervised methods.

3

TRUSTWORTHINESS PREDICTION

Within the field of computational trust, robust statistical prediction methods for determining the future behaviour of a potential trustee have received particular attention. Recalling that trust was defined in Definition 2 as a ' *subjective probability with which an agent [*the truster*] expects that another agent or group of agents [*the trustee*] will perform . . . '*, the framing of trust as a probability estimation problem is intuitive. As such, *trust assessment*, i.e., the process of establishing the subjective probability of Gambetta's definition, is regularly tackled in the related literature. Machine learning and statistical estimation have been the dominant tools applied so far. Traditionally, experience-based Bayes predictors form the basis of many of the proposed trust prediction models, while more recently other prediction paradigms, in parti

In this chapter, the Bayesian foundations of the *CertainTrust* model [174], a state-of-the-art binomial trust model, will be revised and extended. In the first part, Section 3.1, the original version of *CertainTrust* will be augmented by statistically sound methods for a more accurate certainty estimation. To this end, two different approaches will be delineated, a Bayesian Credible Interval Based Certainty Estimator and a Frequentist Confidence Interval Based Certainty Estimator. These new certainty estimators adapt work on binomial proportion confidence intervals from the statistics literature. In order to integrate them with *CertainTrust*, the computation of the *CertainTrust* expectation value is generalised. *CertainTrust* also facilitates the graphical representation of trust and certainty scores via its Human Trust Interface (HTI). The HTI is modified in order to represent certainty scores computed from confidence intervals. This allows the user to graphically interpret the probable dispersion of a trust estimate.

In the latter part of this chapter, Section 3.2, the *CertainTrust* model is extended further to handle multi-categorical input generated from a multinomial process. The newly introduced certainty estimators are extended from binomial-proportion interval-based methods to handle simultaneous certainty estimates for multinomial proportions. Additionally, the HTI representation is adapted for dealing with multinomial opinions.

3.1 BINOMIAL CASE

At its core, trust assessment can be interpreted as a statistical probability estimation problem of determining a probability $P(y)$, where y is a representation of the *particular action* of Gambetta's definition. In

the binomial case, we consider the decision whether or not to trust as a binary classification problem – a truster classifies a trustee as either trustworthy or untrustworthy. In this sense, trustworthiness classification is a discriminatory problem suitably assigned to statistical learning methods. However, in order to satisfy the definition of trust as a subjective probability, assigning a class label is insufficient. Rather, the goal in trust assessment is estimating the *probability of class membership*, establishing just *how* likely a particular trustee is to be trustworthy.

For the binomial case, we will adopt a simplified version of the *CertainTrust* [174] opinion representation, $\omega := (t, c)$. Here, t represents an estimate of the trustworthiness of a particular trustee, while c represents an estimate of how *certain* – that is, *reliable* – the estimate t is.

The probability estimate t will be addressed in its fundamentals in Section 3.1.1. The probability estimation task is addressed as a binomial proportion; this is concordant with the state-of-the-art in computational trust modelling. The final result is identical to the work presented by Ries [174] and similar to many other trust models (see also [113]).

The certainty parameter, c, is an estimate that builds upon the fundamental properties of the probability estimate and the theory of interval estimation for binomial proportions. The work presented in Section 3.1.2 extends the state-of-the-art by providing a statistically sound certainty measure for trust assessment.

3.1.1 *Binomial Probability Estimation*

Placing the process of estimating trustworthiness in the context of probability theory and limiting the scope of the prediction to the immediate future, the aim of trustworthiness prediction is to reliably estimate the probability of the trustee acting in a trustworthy manner in the next interaction with the truster, based upon representative input data. Thus, if $y \in \{0; 1\}$ is the outcome of such a future interaction, the goal is to compute a *conditional* probability $P(y = 1|\mathbf{x})$ given the features $\mathbf{x}$. The set of feature $\mathbf{x}$ serve as input data for the estimation process. For instance, $\mathbf{x}$ can contain a history of prior interactions. Furthermore, for binary outputs, it follows that $P(y = 1|\mathbf{x}) = E(y|\mathbf{x})$. Thus, the binary estimation model entails the computation of a binomial expectation value.

The kind and origin of the representative input data, i.e., $\mathbf{x}$, are important for the construction of estimators that compute a trust estimate. In the following, it is assumed that the trustor can uniquely identitfy the trustee and can evaluate the outcome of an interaction with the trustee after it has taken place. In the binomial case, an interaction can either have a successful outcome, generating a positive

experience, or an unsuccessful one, generating a negative experience. Thus, in terms of Gambetta's definition [64], if the trustee performs the '*... particular action...*' according to the truster's expectations, the interaction will result in a positive outcome; otherwise, a negative outcome will ensue. The outcome of interactions between truster and trustee, which can be thought of as a function of the trustee's performance and the truster's expectations, is typically encoded as 1 for a successful interaction yielding a positive outcome and as 0 for an unsuccessful interaction yielding a negative outcome. Repeated interactions between specific pairs of truster and trustee will therefore yield an *interaction history* of experiences for each pairing, that can be represented as a sample $\{0;1\}^n$, with $n \in \mathbb{N}$ being the number of past interactions[1].

Assuming that unique identity and interaction history is the only information a truster has for evaluating a trustee, trust assessment constitutes a *non-associative* probability estimation problem based on learning the behaviour[2] of the trustee. By *non-associative*, it is meant that no features are observed that would permit the creation of association rules, except for the interaction history. For the time being, we will also not consider *contextual* and *situational* information explicitly, but rather assume that context and situation are implicitly determined beforehand. Thus, they are not parameters of the estimation model per se. Thus, the conditional probability to be established, that is, the probability of success (of a positive outcome), is $P(y = 1|\mathbf{x} \in \{0;1\}^n)$.

In order to facilitate deriving an appropriate estimator of $P(y = 1|\mathbf{x} \in \{0;1\}^n)$ and leverage point estimation techniques (see generally [126]), the sample will be assumed to be generated by a simple stochastic process – a *Bernoulli trial*. Hence, the sample follows a *binomial distribution*. In this model, *concept drift* or *non-stationarity* – that is, a change in the data-generating process – is not accounted for. Detecting and compensating non-stationarity will be discussed as a data-processing step in Section 4.5.

From the assumption of a binomial distribution underlying trust assessment in the binomial case follow a number of conclusions – in particular with regard to the data available at the time of making a trust assessment. These we will leverage in the following. First, from repeated interactions between truster and trustee, an evidence-base of prior experiences between these two specific entities is created. This evidence-base is in the form of a time series of randomly generated values in $\{0;1\}$, so that after n interactions (or *trials*) there exists a specific sample of $\{0;1\}^n$. Second, the probability of success p that

1 The notation $0;1^n$ represents the set of all (ordered) n-tupels consisting of zeros and ones.

2 The fundamental mechanics of this particular kind of learning task are a basic estimation problem, which has been explored exhaustively, for instance, in mathematical learning theory [9, 33].

uniquely describes the binomial distribution, being the *estimand*, is identical for all trials in the time series, and the individual trials are statistically independent.

Thus, a given sample of size n can be considered a random variable X, that follows the binomial distribution with probability mass function (*pmf*) $f(x; p, n) = \binom{n}{x} p^x (1-p)^{n-x}$. Here, x denotes the sum of *successes* in the random variable, that is, the sum of 1-elements in the sample. Accordingly, $n - x$ represents the sum of *failures*, that is, the sum of 0-elements. Because all trials are assumed to be *independent and identically distributed (iid)*, it follows that a trial is distributed according to a Bernoulli distribution with parameter p. The binomial distribution has a single parameter $p \in [0; 1]$ and the data, in the form of a sample $\{0; 1\}^n$, can be summarised in a sufficient statistic, for instance the sum of successes, i.e., $x = \sum_1^n \{0; 1\}^n$.

Under the assumption of *iid* outcomes, the probability of a positive outcome, $P(y = 1 | \{0; 1\}^n)$, is solely dependent on the parameter p of the Bernoulli distribution that is driving the repeated trials. Since p is an unknown parameter, an estimate $\hat{p}$ has to be established from the information available – the sample X. An appropriate estimator δ is, for instance, the sample mean of X, yielding $\hat{p} = \frac{x}{n}$ [126]. The sample mean is the *Maximum Likelihood Estimator (MLE)* for p. Thus, trust assessment in the binomial case is a type of *binomial proportion* estimation problem.

From a Bayesian perspective, it might be desirable to substitute the *posterior Maximum Likelihood Estimator*, which includes a prior. Such a prior can encode subjective, a-priori knowledge and is generally realised as a pseudo-count added to the actually observed interaction history. The choice of prior is discussed in Section 3.1.6, p. 66.

3.1.2 *Binomial Certainty Estimation*

Establishing the probability estimate t of a (simplified) *CertainTrust* opinion $\omega := (t, c)$ as a point estimate $\hat{p} = \frac{x}{n}$ of the binomial proportion of a statistical sample is intuitive, as argued in Section 3.1.1. Even under the assumption of an apparently simple binomial distribution underlying trust assessment in the binomial case, however, estimating the binomial proportion alone is insufficient for reliably determining the true parameter p. Because the point estimate $\hat{p}$ is made from a statistical sample that was generated by a random process, the possibility of *sampling error* has to be taken into account.

The potential for estimating and expressing sampling error has been addressed in various trust models through the notion of modelling *(un-)certainty*, for instance in Jøsang & Ismail [108], Wang & Singh [196], Teacy et al. [189], or in Ries [174].

In the following, we will interpret certainty as an estimate that expresses the reliability of the point estimate $t = \hat{p}$. As such, it is a function of the data contained in the sample.

Definition 6 (Certainty)**.** *Certainty* is an estimate for the reliability of the trust estimate $t = \hat{p}$. Its range is $[0; 1]$. A *Certainty Function* or *Certainty Estimator* is a statistic for computing this estimate.

In this context, *reliability* [154] in a statistical sense is a measure of the variability of a measurement. Reliability accounts for the error contained within the measurement, for instance, sampling error. Determining the *true* reliability of a measurement requires knowing both the true score component and the error component – both of which are – in general – individually unobservable. Thus, a corresponding estimate is required.

Under the given assumptions of a binomially distributed random variable with fixed probability of success p, i.e., generated by a stationary process, the notion of reliability is synonymous with *precision*. Furthermore, because the sample mean is a consistent estimator of the parameter p of the binomial distribution that was used to generate the sample, increased certainty also implies increased *validity/accuracy* of the estimate.

If certainty is understood as an estimate of the precision and accuracy of the point estimate $\hat{p} = t$ with regard to some unobservable parameter p, as it will be here, a large body of work from the field statistical science can be adapted. Through the formalisation of trustworthiness estimation as a binomial probability estimation problem, the application of confidence interval estimation techniques for binomial proportions is enabled. Interval estimation of the probability of success (that is, the parameter p) in a binomial distribution has been one of the *'most basic and important problems in statistical practice'* [27]. It has thus received considerable attention over the past decades.

Conversely, while the concept of (un-)certainty is central to state-of-the-art in Bayesian trust assessment, statistically sound methods of interval estimation have not been considered in the literature on computational trust.

Many of the certainty estimators proposed in the literature [103, 152, 174] are heuristically derived. They leverage the fundamental property that with increasing sample size n the trust estimate $t = \hat{p} = \frac{x}{n}$ converges to the true parameter p of the binomial distribution. This, of course, follows directly from the *strong consistency* of the sample mean as an estimator of the parameter p of the binomial distribution[3], i.e., $\lim_{n\to\infty} \hat{p} = p$.

Recent work by Wang & Singh [196, 197] has contributed considerably to providing improved certainty estimation. They achieved this,

3 Under the assumption given in section 3.1.1

by basing their certainty estimation directly on the prior assumption of the Bayesian estimation process.

3.1.3 *Bayesian Foundations of Certainty Estimation*

Recall that the estimation task at hand is the determination of the value of the parameter p of a binomial distribution with pmf $f(x; p, n) = \binom{n}{x} \cdot p^x \cdot (1-p)^{n-x}$. A random variable X that follows this distribution is assumed to be generated by repeated Bernoulli trials with probability mass function $f(x; p) = p^x \cdot (1-p)^{1-x}$, with $x \in \{0; 1\}$. When expressed as a function of the parameter p, this yields $g(p) \propto p^\alpha \cdot (1-p)^\beta$ for some constants α and β. From this, we can derive a probability distribution of the parameter $p \in [0; 1]$ by multiplying $g(p)$ with an appropriate normalising constant, so that $\int_0^1 g(p)dp = 1$.

This normalisation is achieved by dividing $g(p)$ by the Beta function $B(\alpha, \beta) = \frac{\Gamma(\alpha)\cdot\Gamma(\beta)}{\Gamma(\alpha+\beta)}$, for some $\alpha, \beta > 0$ and $\Gamma(z) = \int_0^\infty t^{z-1} \cdot e^{-t} dt$. The resulting probability distribution is the Beta distribution with the probability density function (pdf) given in Equation 1.

$$f(p; \alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\cdot\Gamma(\beta)} \cdot p^{\alpha-1} \cdot (1-p)^{\beta-1}$$
$$\Gamma(z) = \int_0^\infty t^{z-1} \cdot e^{-t} dt \tag{1}$$

In Bayesian statistics, the Beta distribution is the *conjugate prior distribution* of the Binomial distribution [12, 49], with α and β being positive shape parameters of the distribution. In order to incorporate prior knowledge and since the Beta distribution is undefined for $\alpha, \beta = 0$, a *proper* prior [12] can be guaranteed by adding constant values $\alpha_0, \beta_0 > 0$ to α, β. The resulting Beta distribution with pdf $f(p; \alpha_0, \beta_0)$ represents the *prior (density)* given some prior information about the distribution of p encoded in α_0, β_0.

However, even when no prior information is available, *non-informative priors* can be leveraged to obtain Bayesian estimates. Such a prior should not contain any information about p; that is, it favours no possible values of p over others [12]. Frequently used non-informative priors include the *Uniform prior* with $\alpha_0 = \beta_0 = 1$ for point estimation and the *Jeffreys prior* with $\alpha_0 = \beta_0 = \frac{1}{2}$ for interval estimation [12, 96, 97, 99].

The use of the uniform prior follows from Laplace's *Principle of Insufficient Reason*, by which an equal probability assignment to each point in the parameter space is due to an insufficient reason to suppose an alternative [117, 124]. Reasonably, it is a *reference prior* [117, 99] for the estimation of location parameters.

The non-informative Jeffreys prior is deduced from a *Principle of Invariance* [99] that '*equivalent propositions have the same probability*'. This

requires a rule for determining the prior density in such a manner that it is invariant to the change of variables (for a detailed explanation, see [96]).

One way to satisfy this condition is by choosing the prior density for the unknown parameter p proportional to the square root of the *Fisher Information* [126] – that is, $f(p) \propto \sqrt{\det(\mathbb{I}(p))}$. For the binomial case, we know that $\mathbb{I}(p) = \frac{n}{p \cdot (1-p)}$. Therefore, it follows that

$$\sqrt{\det(\mathbb{I}(p))} = \sqrt{n} \cdot \frac{1}{\sqrt{p} \cdot \sqrt{1-p}}$$

and thus

$$f(p) \propto \frac{1}{\sqrt{p} \cdot \sqrt{1-p}} = p^{-\frac{1}{2}} \cdot (1-p)^{-\frac{1}{2}}$$

Accordingly, $p \sim \mathrm{Beta}(\frac{1}{2}, \frac{1}{2})$ [99].

If instantiated with the sum of successes[4] in sample $X = \{x_1, x_2, \ldots, x_n\}$ of size $n \in \mathbb{N}$, $0 < \alpha = \alpha_0 + \sum_{i=1}^{n} \{x_i, x_i = 1\}$, and the sum of failures in the same sample, $0 < \beta = \beta_0 + \sum_{i=1}^{n} \{x_i, x_i = 0\}$, the Beta distribution gives the probability distribution of the parameter p over the parameter space $[0; 1]$. Because $\int_0^1 f(p; \alpha, \beta) dp = 1$, this is an actual probability distribution of p rather than merely a likelihood function. This instantiation of the Beta distribution constitutes the *posterior probability distribution* of p, conditional on the prior and the evidence contained in the sample X. Following [99], $\mathrm{Posterior\ probability} \propto \mathrm{Prior\ probability} \times \mathrm{Likelihood}$.

In terms of sample mean $\hat{p} \in [0; 1]$ and sample size n, the pdf of the Beta distribution can be reformulated by setting $\alpha = n \cdot \hat{p}$ and $\beta = n \cdot (1 - \hat{p})$, resulting in Equation 2.

$$f(p; \hat{p}, n) = \frac{\Gamma(n)}{\Gamma(n \cdot \hat{p}) \cdot \Gamma(n \cdot (1-\hat{p}))} \cdot p^{n \cdot \hat{p} - 1} \cdot (1-p)^{n \cdot (1-\hat{p}) - 1} \quad (2)$$

The relationship between the binomial distribution and its conjugate prior Beta distribution is harnessed by Bayesian trust models, such as [28, 108, 151, 174], to provide a solid statistical foundation for the computation of trust scores. Here, both the posterior mean, $\frac{\alpha}{\alpha+\beta}$, and the posterior mode[5], $\frac{\alpha-1}{\alpha+\beta-2}$, furnish a consistent estimator for p.

Given the posterior distribution and the definition of certainty (Definition 6, p. 53), we can also provide a certainty estimator built upon the dispersion of the Beta distributed posterior of the parameter p. If we consider the dispersion characteristics of a Beta distributed random variable, for instance measured as variance (Equation 3), it is

4 Note that the sum of successes in a sample, $x = \sum_{i=1}^{n} \{x_i, x_i = 1\}$, is a sufficient statistic for estimating p.

5 Depending on the choice of prior.

governed by two components: the size of the sample, n, and the location of the estimate $\hat{p}$.

$$var[p] = \frac{\alpha \cdot \beta}{(\alpha+\beta)^2 \cdot (\alpha+\beta+1)} = \frac{1}{n+1} \cdot \hat{p} \cdot (1-\hat{p}) \tag{3}$$

This provides an explicit motivation for Jøsang's uncertainty estimator $\frac{1}{n+1}$ [102], which is, in essence, equivalent to a certainty estimator $\frac{n}{n+1}$ that meets the requirements of Definition 6, p. 53. Obviously, $\frac{n}{n+1} \in [0;1[$ and $\lim_{n\to\infty} \frac{n}{n+1} = 1$. This estimator, however, completely disregards the second component of the variance, the location of the estimate $\hat{p}$. Additionally, the certainty estimate $\frac{n}{n+1}$ approaches 1 comparatively quickly, giving comparatively high certainty values for a low number of observed interactions. Thus, the possibility of random sampling error may not be accounted for.

Some recent work on improving certainty estimation [160, 197] has taken into account the impact of the variance of observations on the conditional distribution of p. Wang & Singh [197] term this *conflict in the evidence*, the impact of which they derive somewhat heuristically. However, it follows formally from Definition 3, p. 56, that the dispersion in terms of variance is a parabolic function with $max(var[p]) \propto max(\hat{p} \cdot (1-\hat{p}))$ for fixed n. The maximum is attained at $\hat{p} = \frac{1}{2}$, as can be seen by simple derivation .

Recalling that the dispersion of the Beta distributed posterior gives the conditional dispersion of the possible values of the parameter p, it is clear that a lower dispersion, i.e., a lower variance, results in a higher certainty – they are inversely related. Thus, the variance of the posterior provides a measure of the certainty of a trust estimate. It is, however, not easily interpretable semantically in probabilistic terms. That is, we still desire a certainty estimator that is easy to interpret by semantic standards, but at the same time takes the dispersion properties of the Beta posterior into account.

3.1.4 *Bayesian Interval-Derived Certainty*

In both Bayesian and frequentist interpretations of probability, the reliability of a point estimate is conventionally expressed as an interval estimate of the possible distribution of the estimate. In Bayesian statistic, this interval takes the form of a *credible interval* that can be defined as follows [13]:

Definition 7 (Credible Interval)**.** Let $\pi(\theta;x)$ be a posterior distribution of an unknown parameter $\theta \in \Theta \subseteq \mathbb{R}$ given data x. A $100 \cdot (1-z)\%$ *credible interval* is an interval, so that for some $l, u \in \Theta$ (i.e., $l < u$ in the range of $\pi(\theta;x)$, $[l;u] \subseteq \Theta$), it holds that:

$$1 - z \leqslant P(l \leqslant \theta \leqslant u; x) = \int_l^u \pi(\theta;x)\,d\theta$$

Because the posterior distribution is an actual probability distribution on Θ, with θ a random variable that is conditionally distributed according to $\pi(\theta; x)$, it is possible to speak meaningfully of the probability that $\theta \in [l; u]$ [12]. Thus, we can say that, with a given confidence level $100 \cdot (1 - z)\%$, the true value of the parameter that is being estimated is contained within the credible interval. The value of $z \in [0; 1]$ can be thought of as a level of *residual uncertainty* that a truster is willing to tolerate.

3.1.4.1 *Semantic Interpretation*

In the Bayesian model of inference, the estimand, in our case p, is a random variable with a probability distribution [35]. While this property is leveraged to justify inference over the posterior for determining a trust score via point estimation in various Bayesian trust models, it also furnishes a solid justification for computing a *certainty estimate* from a Bayesian interval estimate. Additionally, the extension of well-established estimation methodologies contributes to the semantic interpretability of certainty estimators that are based upon Bayesian credible intervals. By basing the certainty estimate on a well-established measure of the possible dispersion of the estimand parameter, such as defining certainty as the length of a specific confidence/-credibility interval, the certainty estimate can be readily related to a standard measurement of statistical variability.

The reliability of an inference on the unimodal (for $\alpha, \beta > 1$) Beta posterior, with regard to p, is expressed based on the statistical dispersion of the random variable p. A statement about the certainty, $c \in [0; 1]$, of a parameter estimate $\hat{p}$ of p should be a probabilistically interpretable measure of just this dispersion. From the definition of the credible interval (Definition 7, p. 56), we have three constants: the lower bound of the interval, l, the upper bound of the interval, u, and the desired confidence level $100 \cdot (1 - z)\%$. The bounds $l \in [0; 1]$ and $u \in [0; 1]$, with $l \leqslant u$, enclose an interval that can be said – because it has a Bayesian derivation – to contain the random parameter $p \in [0; 1]$ with a probability of at least $100 \cdot (1 - z)\%$. That is, the probability that $p \notin [l, u]$ is smaller than $100 \cdot z\%$.

However, the credible interval is characterised by a tuple $(l, u) \in [0, 1]^2$, instead of the scalar certainty estimate that is desired when using the *CertainTrust* opinion representation $\omega := (t, c, f)$. In survey sampling, a statistic usually used to express random sampling error in a scalar is the *margin-of-error*, defined as "*the half-width of the confidence interval*"[6] for a given estimate [130], i.e., for $\hat{p}$. The confidence intervals used in survey statistics are typically constructed using the normal approximation derived from the Wald large sample

6 Although there is a difference of interpretation between a frequentist confidence interval and a Bayesian credible interval, assume them to be equivalent for now. The difference will be briefly addressed in Section 3.1.5, p. 62

test for the binomial case [27]. Since this approximation supposes a normal distribution of the parameter p over $[0;1]$, the margin-of-error in this case, $\kappa_{\alpha/2} \cdot \sqrt{\frac{1}{n} \cdot \hat{p} \cdot (1-\hat{p})}$, is symmetrical around $\hat{p}$ ($\kappa_{\alpha/2}$ is the $100 \cdot (1-\frac{z}{2})$ percentile of the standard normal distribution).

The Wald approximation in the binomial case is very inaccurate [27]. The Bayesian credible interval with a Beta prior is more accurate, in terms of coverage properties [27]; the resulting Beta posteriors are, generally, not symmetric. Additionally, we are not interested in the actual endpoints of the interval, but in a statistic that describes the reliability of the estimate $\hat{p}$. As such, the *length of the credible interval* suffices.

The length of the credible interval is contained in $[0;1]$ and approaches zero with increasing sample size n, as is shown below. Conveniently, the interval length is given in the same scale as p – that is, in percentage points if multiplied by 100. We know therefore that, with a residual uncertainty of $100 \cdot z\%$, the likely dispersion of $\hat{p}$ is within a range of percentage points given by the interval length[7].

What remains is to actually construct a credible interval for $p \sim \mathrm{Beta}(\alpha, \beta)$.

3.1.4.2 *Interval Construction*

Obviously, for a given value of z, there is no *unique* credible interval. A usual approach for attaining a specific credible interval for θ would be to compute the interval that has the minimal length. This can be achieved by considering the interval which has the highest posterior density [12, 13], the Bayesian HPD interval. However, HPD intervals are considerably harder to compute than and do not perform as well as *equal-tailed* credible intervals in frequentist terms [27]. Since many natural conjugate priors, in particular the Beta prior distributions we are interested in here, are unimodal for $\alpha, \beta > 1$, the resulting posteriors are unimodal as well. From this, the use of intervals, instead of more general credibility sets [12, 13], results. Furthermore, since we are considering a binomial proportion problem that is necessarily unimodal and has a finite range $[0;1]$, equal-tailed intervals approximate the HPD intervals well [12].

Definition 8 (Equal-tailed Credible Interval)**.** A credible interval is *equal-tailed* if the probability to exclude θ from $[l;u]$ is the same for both the lower bound l and the upper bound u:

$$\frac{z}{2} \leqslant P(\theta < l; x) = P(\theta > u; x)$$

Applied to the binomial proportion estimation problem that is constituted by trust assessment, we can thus formulate a credible interval for the estimate of the parameter p using the non-informative Jeffreys prior $\mathrm{Beta}(\frac{1}{2}, \frac{1}{2})$. Following [27], we can define:

7 By Definition 7, p. 56, the trust estimate $\hat{p}$ is contained within the credible interval.

Definition 9 (Jeffreys Interval)**.** Let $X \sim \mathrm{Bin}(n, p)$ and p have a non-informative prior distribution $\mathrm{Beta}(\frac{1}{2}, \frac{1}{2})$. Furthermore, let $B(z; \alpha, \beta)$ denote the z quantile of a $\mathrm{Beta}(\alpha, \beta)$ distribution. Then, the equal-tailed $100 \cdot (1-z)\%$ *Jeffreys (prior) interval* is defined as

$$CI_J = [L_J(x), U_J(x)]$$

where $L_J(0) = 0$ and $U_J(n) = 1$ and otherwise

$$L_J(x) = B(\frac{z}{2}; x + \frac{1}{2}, n - x + \frac{1}{2})$$

and

$$U_J(x) = B(1 - \frac{z}{2}; x + \frac{1}{2}, n - x + \frac{1}{2})$$

The interval is the central $1-z$ posterior probability interval that omits $\frac{z}{2}$ in each tail. It is modified for the special cases of $x = 0$ and $x = n$ by adjusting the bounds according to Definition 9, i.e., $L_J(0) = 0$ and $U_J(n) = 1$. This is done in order to guarantee good frequentist performance, with regard to frequentist *coverage probabilities* [12].

Based on the Jeffreys interval, we can now define a Bayesian Certainty Estimator by harnessing that $\lim_{n\to\infty} \mathrm{length}(CI_J) = 0$ and that $\hat{p}$ is a consistent estimator, $\lim_{n\to\infty} \hat{p} = p$.

Definition 10 (Credibility Interval-based Certainty Estimator)**.** The *Credibility Interval-based Certainty Estimator* for a trust estimate $t = \hat{p} = \frac{x}{n}$ and an *acceptable residual uncertainty level* (confidence level) $(100 \cdot z)\%$, is defined as

$$C_{J;(100\cdot z)\%}(x, n) := 1 - (U_J(x) - L_J(x))$$

This estimator provides a statistically sound estimate of dispersion of the estimand parameter p. It is derived directly from the Beta distribution, and thus conforms to the Bayesian interpretation of trust assessment.

3.1.4.3 *Monotonicity Property for fixed $\hat{p}$, $n \to \infty$*

Wang & Singh [197] postulate three properties that they consider important for certainty estimators. The first two of those are concerned with monotonicity of the certainty functions under different paramaterizations. First, they demand that for a fixed proportion estimate $\hat{p}$, the certainty estimate should increase with sample size n for $n > 0$. This property follows directly from the construction of the Credibility Interval-based Certainty Estimator.

Property 1. *Fix the proportion $\hat{p} = \frac{x}{n}$. Then $C_{J;(100\cdot z)\%}(x, n)$ increases with n for $n > 0$.*

PROOF Recall that the Beta distribution is unimodal for $\alpha, \beta \geqslant 1$, and that $C_{J;(100\cdot z)\%}$ equals $1 - \mathtt{length}(CI_J)$ for some confidence level z. Furthermore, CI_J is an equal-tailed, central credible interval for p. It follows, that for a fixed proportion estimate $\hat{p}$, $C_{J;(100\cdot z)\%}$ is proportional to the standard deviation $\mathtt{sd}[p] = \sqrt{\mathtt{var}[p]}$; according to Equation 3 we can thus rewrite: $1 - C_{J;(100\cdot z)\%} \propto \sqrt{\frac{1}{n+1} \cdot \hat{p} \cdot (1-\hat{p})}$. As $\hat{p} \cdot (1-\hat{p})$ is assumed to be fixed, this can be simplified to $1 - C_{J;(100\cdot z)\%} \propto \sqrt{\frac{1}{n+1}}$. As $\lim_{n\to\infty} \sqrt{\frac{1}{n+1}} = 0$ and $\sqrt{\frac{1}{n+1}}$ is strictly monotonically decreasing, Wang & Singh's first property follows.

3.1.4.4 *Monotonicity Property for fixed n, variable $\hat{p} \in [0;1]$*

Wang & Singh's [197] second property demands that a certainty estimator for a binomial proportion trust estimation problem should report minimum certainty at $\hat{p} = 0.5$ for fixed sample size n. Furthermore, the certainty estimate should be monotonically decreasing with increasing $\hat{p}$ for $\hat{p} \in [0;0.5]$ and monotonically increasing with increasing $\hat{p}$ for $\hat{p} \in [0.5;1]$. Again, this property follows directly from the construction of the Credibility Interval-based Certainty Estimator.

Property 2. *For fixed sample size n and variable proportion $\hat{p} = \frac{x}{n}$, $C_{J;(100\cdot z)\%}(x,n)$ is decreasing when $0 \leqslant \hat{p} \leqslant \frac{1}{2}$, and increasing when $\frac{1}{2} \leqslant \hat{p} \leqslant 1$. For fixed n and variable $\hat{p}$, $C_{J;(100\cdot z)\%}(x,n)$ is minimised at $\hat{p} = 0.5$.*

PROOF Since $1 - C_{J;(100\cdot z)\%} \propto \sqrt{\frac{1}{n+1} \cdot \hat{p} \cdot (1-\hat{p})}$, as outlined above, when we fix n this time, we know that $1 - C_{J;(100\cdot z)\%} \propto \sqrt{\hat{p} \cdot (1-\hat{p})}$ for variable $\hat{p}$. Straightforward derivation of $\hat{p}^{\frac{1}{2}} \cdot (1-\hat{p})^{\frac{1}{2}}$ in $\hat{p}$, shows that this function is maximised at $\hat{p} = 0.5$. Furthermore, since $\hat{p} \cdot (1-\hat{p})$ is monotonically increasing with increasing $\hat{p}$ for $\hat{p} \in [0;0.5]$ and monotonically decreasing with increasing $\hat{p}$ for $\hat{p} \in [0.5;1]$, Wang & Singh's second property follows.

3.1.4.5 *Bijection to Evidence Space*

The third property demanded by Wang & Singh [197] is a bijection between the opinion space, in our case given by *CertainTrust* opinions $\omega := (t,c,f)$, and the (simplified) evidence space $E = \{(r,s) | r \geqslant 0, s \geqslant 0, n = r+s\}$, as modelled by [103, 197]. The evidence representation (r,s) is a simple transformation from the representation more commonly found in the statistics literature, e.g., [12, 27], (x,n), where x is the sum of successes and n the sample size. Thus, $x = r$ and $s = n - x$; that is, r is the sum of successes and s is the sum of failures in a sample of size $n = r + s$. Note that x, respectively r, is a *sufficient statistic* for sample $X \sim \mathtt{Bin}(n,p)$ [35].

For the moment, we will be omitting the initial trust parameter f of the CertainTrust opinion space, a parameter that is typically

assigned by the user and expresses subjectivity. The prior distribution induced by the parameter f will be discussed in Section 3.1.6. Since the initial prior does not impact the likelihood-only posterior distribution, it does not contribute directly to the computation of t and c in the *CertainTrust* model and can be considered a system parameter. Thus, assuming f as constant, the relation Z, from evidence space to opinion space, is as given above. Hence, $\omega := (t, c, f) = (\frac{x}{n}, C_{J;(100\cdot z)\%}(x, n), f) = (\frac{r}{r+s}, C_{J;(100\cdot z)\%}(r, r+s), f)$; for $r + s = 0$, we follow Ries [173] and posit $t = \frac{1}{2}$ in this case.

Property 3. *There exists a* bijection *from the* evidence space $E = \{(r, s) | r \geqslant 0, s \geqslant 0, n = r + s\} = \{(x, n) | n \geqslant x \geqslant 0, x = r, n = r + s\}$ *to* opinion space $O = \{(t, c) | 0 \leqslant t \leqslant 1, 0 \leqslant c \leqslant 1\}$, *such that* $Z(x, n) = (t, c)$ *and* $Z^{-1}(t, c) = (x, n)$, *where* $t = \frac{x}{n}$ *and* $c = C_{J;(100\cdot z)\%}(x, n)$.

PROOF That Z is bijective can be shown by applying the same method as [197]. Since $t = \frac{x}{n}$, we only need to show that we can uniquely determine n from $C_{J;(100\cdot z)\%}(x, n)$. Following the arguments of [197], the existence and uniqueness of n is proved by showing that

1. $C_{J;(100\cdot z)\%}(x, n)$ is monotonically increasing for fixed $\hat{p} = t = \frac{x}{n}$ and $n > 0$ (Property 1)
2. $\lim_{n\to\infty} C_{J;(100\cdot z)\%}(x, n) = 1$ for fixed $\hat{p} = t = \frac{x}{n}$
3. $\lim_{n\to 0} C_{J;(100\cdot z)\%}(x, n) = 0$ for fixed $\hat{p} = t = \frac{x}{n}$

The first point follows from Property 1. For the second and third point, the behaviour of $C_{J;(100\cdot z)\%}(x, n)$ in the limits towards zero and infinity can be easily shown by harnessing that for fixed $\hat{p}$ it holds that $1 - C_{J;(100\cdot z)\%}(x, n) \propto \sqrt{\frac{1}{n+1}}$.

The function for computing $C_{J;(100\cdot z)\%}(x, n)$ is open form. Thus, the inverse function needed to compute Z^{-1} is also unavailable in closed form representation. Since Wang & Singh's approach [197] has the same restrictions, they propose a binary search algorithm (Algorithm 1, p. 62) of complexity $\Omega(-\lg \epsilon)$, which we will adopt[8].

The upper bound n_{max} can be determined by exponentially increasing n and computing $C_{J;(100\cdot z)\%}(\hat{p} \cdot n, n)$ until a value is found for which the certainty estimate, c, that is recorded in the opinion $\omega := (t, c, f)$ is exceeded [197]. Alternatively, an upper bound may be established via look-up in pre-computed tables for various important or frequently used certainty values and confidence levels, such as $c = 0.99999$ ("*five nines*") at a confidence level of $z = 0.95$ (leveraging Property 1). These tables can be held reasonably short by providing only worst case estimates; that is, only maintaining table entries for $p = \frac{1}{2}$ (Property 2).

8 For small $n \in \mathbb{N}$, approximate solutions for $C_{J;(100\cdot z)\%}(x, n)$ and its inverse may be provided in pre-computed tables, for instance given in Appendix B.

Data: Trust estimate $t = \hat{p} = \frac{x}{n}$, certainty estimate $C_{J;(100\cdot z)\%}(x, n)$
Result: Sample size n, and sum of successes x

// Initialize parameters
t=$\hat{p}$;
c=$C_{J;(100\cdot z)\%}$;
$n_1 = 0$;
$n_2 = n_{max}$;
// Approximate n to specified precision ϵ
while $n_2 - n_1 \geqslant \epsilon$ **do**
    $n = \frac{n_1+n_2}{2}$;
    if $C_{J;(100\cdot z)\%}(t \cdot n, n) < c$ **then**
        $n_1 = n$
    end
    else
        $n_2 = n$
    end
end
return $n, x = t \cdot n$

Algorithm 1: Calculation of $(x, n) = Z^{-1}(t, c)$ [197]

3.1.5 *Confidence Interval-Derived Certainty*

The method of deriving a certainty estimate from Bayesian intervals presented in the previous section (Section 3.1.4.2, in particular Definitions 8 to 10) leverages statistically sound principles of interval estimation. Inconveniently, however, it is unavailable in a closed form representation – a shortcoming shared by Teacy's [189] and Wang & Singh's methods [197].

A closed form for both the certainty estimator $C(x, n)$, as well as its inverse function, offers quicker comprehension and a more compact representation of the mathematical operations used for computing the estimate. At the same time, the easy interpretability of the credibility interval based certainty estimator should be retained.

Recall that the certainty estimator $C_{J;(100\cdot z)\%}(x, n)$ is based on an open form interval estimate, the $(100 \cdot z)\%$ credibility interval of the Jeffreys prior. The aim thus is to find a closed form way of computing an alternative interval. This interval should approximate the behaviour of the Jeffreys interval.

Recall that, in the presented form, trust assessment is essentially a binomial proportion estimation problem. As such, the goal is to estimate a single parameter, p, the probability of success from a binomial distribution. Furthermore, the data in the sample $X \in \{0; 1\}^n$ can be summarised in a single sufficient statistic [35], $x = \sum_1^n \{0; 1\}^n$.

In this very specific case, i.e., single parameter estimation and a summarising sufficient statistic, the frequentist *confidence* and the Bayes-

ian *credible* interval are highly similar[9] [12, 97]. Thus, the – normally carefully maintained – difference in interpretation and meaning between frequentist and Bayesian method (see generally [12, 35]) does not impact the semantic interpretability of the certainty estimate if we substitute a confidence interval for the credible interval of similar frequentist performance.

The good frequentist performance of the Jeffreys prior interval has been remarked upon in the literature [27, 117, 199]. Together with two other interval estimators, the *Wilson interval* [204] and the *Agresti-Coul interval* [4], it has been recommended for the interval estimation of binomial proportions [27]. The Jeffreys and Wilson intervals, in particular, perform similarly well for small sample sizes. Therefore, basing a certainty estimator on the Wilson interval should yield similar performance, while overcoming the Jeffrey interval's limitation of not being available in closed from.

3.1.5.1 *Interval Construction*

Because of its good performance – in terms of expected length and coverage probabilities [27] – as well as a reasonably compact closed form representation, the Wilson interval lends itself to computing a confidence interval-derived certainty estimate.

Definition 11 (Wilson Interval)**.** Let $\Phi(z)$ be the standard normal distribution function and κ the $100 \cdot (1 - \frac{z}{2})$ percentile of the standard normal distribtution, i.e., $\kappa = \Phi^{-1}(1 - \frac{z}{2})$. $\hat{p} = \frac{x}{n}$, $X \sim \mathrm{Bin}(n, p)$. The $100 \cdot (1 - z)\%$ *Wilson interval* [27, 204] is defined as:

$$\mathrm{CI}_W = \frac{x + \frac{\kappa^2}{2}}{n + \kappa^2} \pm \frac{\kappa \cdot \sqrt{n}}{n + \kappa^2} \cdot \sqrt{\hat{p} \cdot (1 - \hat{p}) + \frac{\kappa^2}{4 \cdot n}}$$

Thus, the upper bound of the Wilson confidence interval, $U_W(x)$, and the corresponding lower bound, $L_W(x)$, for a sample $X \sim \mathrm{Bin}(n, p)$ of size $n \in \mathbb{N}$, are given by:

$$U_W(x) = \frac{x + \frac{\kappa^2}{2}}{n + \kappa^2} + \frac{\kappa \cdot \sqrt{n}}{n + \kappa^2} \cdot \sqrt{\hat{p} \cdot (1 - \hat{p}) + \frac{\kappa^2}{4 \cdot n}}$$

and

$$L_W(x) = \frac{x + \frac{\kappa^2}{2}}{n + \kappa^2} - \frac{\kappa \cdot \sqrt{n}}{n + \kappa^2} \cdot \sqrt{\hat{p} \cdot (1 - \hat{p}) + \frac{\kappa^2}{4 \cdot n}}$$

Analogously to the Credibility Interval-based Certainty Estimator, Definition 10, p. 59, the Confidence Interval-based Certainty Estimator is given in the following Definition 12:

9 The similarity, however, does not generally extend beyond this special case. In fact, inferences drawn from confidence and credible intervals can differ considerably from each other [35].

Definition 12 (Confidence Interval-based Certainty Estimator)**.** The *Confidence Interval-based Certainty Estimator* for a trust estimate $t = \hat{p} = \frac{x}{n}$ and an *acceptable residual uncertainty level* (confidence level) $(100 \cdot z)\%$, is defined as

$$\begin{aligned} C_{W;(100\cdot z)\%}(x, n) &:= 1 - (U_W(x) - L_W(x)) \\ &= 1 - \left(2 \cdot \frac{\kappa \cdot \sqrt{n}}{n + \kappa^2} \cdot \sqrt{\hat{p} \cdot (1 - \hat{p}) + \frac{\kappa^2}{4 \cdot n}}\right) \end{aligned}$$

κ is the $100 \cdot (1 - \frac{z}{2})$ percentile of the standard normal distribution.

The desired properties for certainty estimators postulated by Wang & Singh [197] hold for the Confidence Interval-based Certainty Estimator. The procedure for showing these properties is similarly straightforward as for the Credibility Interval-based Certainty Estimator, owing to the construction of the Wilson interval by inverting the null standard error test statistic, $\sqrt{\frac{p \cdot (1-p)}{n}}$ [27, 204]. Additionally, the bijection between opinion and evidence space can now be represented in a closed form.

3.1.5.2 *Monotonicity Property for fixed* $\hat{p}$, $n \to \infty$

Property 1. *Fix the proportion* $\hat{p} = \frac{x}{n}$*. Then* $C_{W;(100\cdot z)\%}(x, n)$ *increases with* n *for* $n > 0$*.*

Wang & Singh's [197] first demand is that for a fixed proportion estimate $\hat{p}$, the certainty estimate should increase with sample size n for $n > 0$.

PROOF This property follows from the behaviour of $\frac{\sqrt{n}}{n}$ and $\sqrt{\frac{1}{n}}$ in the limit, both being strictly monotonically decreasing. That is, $\lim_{n\to\infty} \frac{\sqrt{n}}{n} = 0$ and $\lim_{n\to\infty} \sqrt{\frac{1}{n}} = 0$, from which, for fixed $\hat{p}$, it follows that:

$$\lim_{n\to\infty} \left(2 \cdot \frac{\kappa \cdot \sqrt{n}}{n + \kappa^2} \cdot \sqrt{\hat{p} \cdot (1 - \hat{p}) + \frac{\kappa^2}{4 \cdot n}}\right) = 0$$

Since the subtrahend of the Confidence Interval-based Certainty Estimator, Definition 12, p. 64, is strictly monotonically decreasing for increasing n and is bounded by 0, with $\hat{p}$ and κ as constants, it follows that $C_{W;(100\cdot z)\%}(x, n)$ is strictly monotonically increasing with an upper bound at 1.

3.1.5.3 *Monotonicity Property for fixed* n*, variable* $\hat{p} \in [0; 1]$

Wang & Singh's [197] second property demands that a certainty estimator for a binomial proportion trust estimation problem should report minimum certainty at $\hat{p} = 0.5$ for fixed sample size n. Furthermore, the certainty estimate should be monotonically decreasing with

increasing $\hat{p}$ for $\hat{p} \in [0; 0.5]$ and monotonically increasing with increasing $\hat{p}$ for $\hat{p} \in [0.5; 1]$. The argument for the Confidence Interval-based Certainty Estimator's compliance to this property is identical to that put forth for the Credibility Interval-based Certainty Estimator.

Property 2. *For fixed sample size* n *and variable proportion* $\hat{p} = \frac{x}{n}$, $C_{W;(100\cdot z)\%}(x, n)$ *is decreasing when* $0 \leqslant \hat{p} \leqslant \frac{1}{2}$, *and increasing when* $\frac{1}{2} \leqslant \hat{p} \leqslant 1$. *For fixed* n *and variable* $\hat{p}$, $C_{W;(100\cdot z)\%}(x, n)$ *is minimised at* $\hat{p} = 0.5$.

PROOF When we fix n, we know that $1 - C_{W;(100\cdot z)\%} \propto \sqrt{\hat{p} \cdot (1 - \hat{p}) + K}$ for variable $\hat{p}$ and constant $K = \frac{\kappa^2}{4 \cdot n} > 0$. Straightforward derivation of $\sqrt{\hat{p} \cdot (1 - \hat{p}) + K}$ in $\hat{p}$, shows that this function is maximised at $\hat{p} = 0.5$. Furthermore, since $\hat{p} \cdot (1 - \hat{p})$ is monotonically increasing with increasing $\hat{p}$ for $\hat{p} \in [0; 0.5]$ and monotonically decreasing with increasing $\hat{p}$ for $\hat{p} \in [0.5; 1]$, Wang & Singh's second property follows.

3.1.5.4 *Bijection to Evidence Space*

The existence of a bijective relation Z between evidence and opinion space and the uniqueness of n, when using the Confidence Interval-based Certainty Estimator $C_{W;(100\cdot z)\%}(x, n)$, follows from applying exactly the same steps as already shown in Section 3.1.4.5, p. 60.

Whereas the inverse relation from opinion to evidence space Z^{-1} was only given in algorithmic form for the Credibility Interval-based Certainty Estimator (Algorithm 1, p. 62), the closed form of the Confidence Interval-based Certainty Estimator permits the representation of Z^{-1} in a closed form as well. The inverse relation $Z^{-1}(t, c) = (x, n)$ is given in the following Definition 13.

Definition 13 (Inverse Confidence Interval-based Certainty)**.** Let a *CertainTrust* opinion tuple, $\omega = (t, c)$, with $t = \hat{p} = \frac{x}{n}$ and $c = C_{W;(100\cdot z)\%}(x, n)$, be given. Furthermore, let $(100 \cdot z)\%$, the acceptable residual uncertainty level under which the certainty estimate c was computed, be known and correspondingly, let κ be the $100 \cdot (1 - \frac{z}{2})$ percentile of the standard normal distribution.

The relation $Z^{-1}(t, c) = (x, n)$ is given by computing:

$$n = \frac{-\kappa^2 \cdot \left(2u^2 - 4\hat{p} + 4\hat{p}^2\right) + \sqrt{4u^2 \cdot \kappa^4 \cdot (1 - u^2) + \kappa^4 \cdot (2u^2 - 4\hat{p} + 4\hat{p}^2)^2}}{2u^2}$$

$$x = \hat{p} \cdot n$$

with $u = 1 - c$ (i.e., the length of the Wilson interval).

In Definition 13, the inverse relation $Z^{-1}(t, c) = (x, n)$ is constructed by solving $2 \cdot \frac{\kappa \cdot \sqrt{n}}{n + \kappa^2} \cdot \sqrt{\hat{p} \cdot (1 - \hat{p}) + \frac{\kappa^2}{4 \cdot n}} = 1 - c$, the length of the Wilson interval, for n. Since $t = \hat{p} = \frac{x}{n}$ is known, x is trivial to compute, once n is known: $x = \frac{x}{n} \cdot n$.

The confidence interval derived certainty statistic based on the Wilson interval provides a reasonably compact closed form that can be efficiently computed. This holds for both the certainty estimator and its inverse.

3.1.6 *Initial Trust Value*

The initial trust and weight parameters of *CertainTrust* [174], f and w respectively, determine the Bayesian prior distribution. As has been stated before, trust assessment with binary inputs constitutes a binomial proportion estimation problem. Consequently, this justifies the use of the Beta distribution as a conjugate prior for estimating the parameter $p \in [0;1]$ of a binomial distribution.

By the definition of conjugate distributions [167], a conjugate prior is in the same family of distributions as the posterior. The posterior for a binomial proportion estimation problem has already been shown to be distributed according to a Beta distribution. In order to instantiate the Beta distributed prior from *CertainTrust* parameters, a straightforward bijection between the parameters of a Beta distribution (Equation 1, p. 54), α and β, and the corresponding parameters, f and w of *CertainTrust* is required. This bijection is given by using the alternative[10] representation of the Beta distribution in Equation 2, p. 55 with parameters f and w substituted for $\hat{p}$ and n:

Definition 14 (Initial instantiation of Beta prior with *CertainTrust* parameters)**.** The Beta prior, $g(p;\hat{p},n)$, for a binomial trust estimation problem is instantiated from *CertainTrust* parameters f and w in the following manner:

$$\begin{aligned} \hat{p} &= f, \\ n &= 2 \cdot w, \\ g(p;\hat{p},n) &= \frac{\Gamma(n)}{\Gamma(n\cdot\hat{p})\cdot\Gamma(n\cdot(1-\hat{p}))} \cdot p^{n\cdot\hat{p}-1} \cdot (1-p)^{n\cdot(1-\hat{p})-1} \end{aligned}$$

The *CertainTrust* parameters f and w determine the shape of the prior Beta distribution. Technically, $f \in [0;1]$ encodes a binomial proportion, while $2 \cdot w \in \mathbb{R}^+$ represents a *pseudo count* of subjective experiences that is partitioned according to f. Their concrete choice determines whether or not the resulting prior distribution is an *informative* or a *non-informative* prior.

INFORMATIVE PRIORS Bayesian statistics allows the integration of *subjectivism* into the statistical estimation process by including personal or subjective information in the prior distribution [12, 98]. Informative priors encode (subjective) *a priori* knowledge about the dis-

10 Reformulated to be parameterised with parameters n (sample size) and $\hat{p}$ (proportion of successes), as frequently encountered in the statistics literature (e.g., [27]).

tribution of the estimand parameter p – beyond the information contained in the analysed sample X. For the binomial case, initial trust values $f \neq \frac{1}{2}$, or large weight values w encode presuppositions on the distribution of the estimand. For $f \neq \frac{1}{2}$ it is obvious that the assumed prior distribution of *successes* and *failures* in a Bernoulli process is not identical, therefore favouring one or the other and encoding *a priori information*. For $w > 1$, more weight is given to a particular value of f. That is, the prior is given additional importance in comparison to the likelihood part of the posterior.

In computational trust, informative priors can encode – but is not limited to – aspects of systemic or institutional trust [142]. They can play an important role in adding information to the trustworthiness estimation process that are not quantified by past interactions between truster and trustee. This can include factual knowledge of the truster about the trustee that stems from familiarity or social relations, but also from other types of trustworthiness prediction, such as stereotyping approaches, such as the one presented in Chapter 5.2.

NON-INFORMATIVE PRIORS Non-informative priors were already briefly discussed in the previous section on certainty estimation (Section 3.1.2), with a prominent one, the Jeffreys prior, used for the construction of a certainty statistic. In Bayesian statistical practice, non-informative priors form the basis of most Bayesian analyses and are constructed according to formal rules [117]. Rather than include subjective *a priori* information, the choice of a non-informative prior is governed by the desire for *objectivism* in Bayesian analysis [98]. In other words, a non-informative prior models the analyst's *ignorance*.

This, however, still leaves the question of which actual non-informative *reference* prior to choose – that is, how to instantiate the conjugate prior Beta distribution to represent ignorance.

The most prevalent interpretation of reference priors among statisticians [117] holds that '*... there is no unique prior that represents ignorance*', but rather that reference priors are chosen by public agreement as a default under insufficient information. Three particular different reference priors are frequently encountered in the given setting of binomial proportion estimation.

- *Uniform Prior*: $\mathtt{Beta}(1,1) \leftrightarrow f = \frac{1}{2}, w = 1$. As the reference prior most often chosen in the related work on Bayesian trust modelling , it forms the *default prior* in a number of trust models (e.g., [108, 174, 189, 197]). It presupposes a flat prior distribution[11] of the estimand p over its parameter space $[0;1]$. The Uniform distribution is the original prior deduced by Laplace via the *Principle of Insufficient Reason* [124].

11 For a discussion of the uniformity assumptions underlying the *Rule of Succession* [124] and the Uniform prior, see [186].

- *Haldane's Prior*: $\mathtt{Beta}(0,0) \leftrightarrow f = \frac{1}{2}, w = 0$. The use of a Uniform prior has been disputed by a number of authors (e.g., Jaynes [96] and Jeffreys [99]), because it does not *'... seem to correspond well with the inductive reasoning which we all carry out intuitively'* [96]. In order to address this issue, Jaynes proposes the use of Haldane's prior, derived on the basis of transformation groups, that describes a state of complete ignorance[12], ameliorating the non-intuitive effects of the *Rule of Succession*. However, this prior is *improper* – that is, it does not integrate. This impropriety can yield an improper posterior (in case $x = 0$ or $x = n$), hampering Bayesian analysis. In case of a proper posterior, i.e., $x > 0$ and $x < n$, the resulting Bayesian Posterior Maximum Likelihood estimate coincides with the frequentist Maximum Likelihood estimate.

- *Jeffreys Prior*: $\mathtt{Beta}(\frac{1}{2}, \frac{1}{2}) \leftrightarrow f = \frac{1}{2}, w = \frac{1}{2}$. Jeffreys prior, the principle behind which has already been delineated in Section 3.1.3, p. 55, models ignorance through reduction of the Fisher information [99]. By shifting probability mass towards the end points of the parameter space of p, it seeks to address the non-intuitive behaviour of the *Rule of Succession*, while guaranteeing a proper posterior. In the binomial case, i.e., $\mathtt{Beta}(\frac{1}{2}, \frac{1}{2})$, Jeffreys prior follows an Arcsine function, $f(x) = (\pi \cdot \sqrt{x \cdot (1-x)})^{-1}$ for $x \in [0; 1]$.

The Uniform and Haldane's priors have both been advocated for use in trust models, in particular for establishing the point estimate of the trustee's trustworthiness. For instance, Jøsang & Ismail [108] use the Uniform distribution as a default prior instantiation in their Beta Reputation System, while Aberer & Despotovic [42], through their use of a Maximum Likelihood Estimator (*MLE*), implicitly suppose Haldane's prior. For *CertainTrust*, Ries [174] applies both Haldane's prior and the Uniform prior. The former is, again implicitly, assumed in the *MLE* $t = \hat{p} = \frac{x}{n}$. The latter is used as a default instantiation when computing the final expectation value on the trustworthiness score, $E = c \cdot t + (1-c) \cdot f$, with default parameters $f = \frac{1}{2}$ and $w = 2$, in case $c = 0$.

Jeffreys prior is not readily encountered in the related work on computational trust. However, its main application in Bayesian statistics is in interval estimation – as leveraged in the construction of a credible-interval derived certainty estimator in Section 3.1.4, p. 56. In fact, for the point estimation of the single parameter p of a binomially distributed random variable, Jeffreys himself proposed the use of the Uniform prior [99].

12 Jaynes' argument [96] is that the Uniform prior already contains information – in particular, that both outcomes, i.e., success and failure, of a Bernoulli experiment are *actually* possible.

The actual choice of non-informative prior distribution is a matter of convention – in computational trust, as in the wider field of Bayesian statistics [117]. The Uniform prior is biased towards $\frac{1}{2}$, and consequently – in the mean – less accurate than the *MLE* (i.e., using Haldane's prior in a Bayesian setting), under the synthetic assumption of a sample $X \sim \mathtt{Bin}(n, p)$, with a stable, unchanging parameter p. However, the assumption $X \sim \mathtt{Bin}(n, p)$ is *not* put in place because it accurately models trustee behaviour in computational trust. Rather, it is an assumption made for *statistical convenience*. It implies independent and identically distributed (*iid*) or exchangeable random variables, an assumption that enables a straightforward and statistically sound estimation process and justifies the prediction of future events based on past experience.

While it can be shown that the *MLE* is the most accurate estimator under the assumption of $X \sim \mathtt{Bin}(n, p)$ [181] – which is highly unsurprising, due to the very nature of the *MLE* under stationarity – the choice of prior should reflect *real world* instead of theoretical performance. By biasing the estimation, the Uniform prior trades robustness for accuracy. Whether or not this is warranted, is a matter that depends on the application scenario and necessitates an analysis of the utility of the tradeoff between robustness and accuracy in the real world application[13]. For a more theoretic discussion of the performance of various priors used for estimation, see, for instance, [40].

3.1.7 *Adjusted Expectation Value Computation*

By changing the semantic interpretation of the certainty value c of a *CertainTrust* opinion $\omega := (t, c, f)$ to a dispersion-based reliability statistic, i.e., the length of specific credible interval, the computation of the *CertainTrust* expectation value has to be adjusted. In *CertainTrust* [173], the trust score estimate t represents the frequentist *MLE*, $\frac{x}{n}$. Under the absence of frequency information, i.e., $n = 0 \leftrightarrow c = 0$,the expectation value E represents the Bayesian *posterior MLE*, $\frac{x+1}{n+2}$, with a Uniform prior. Since $E = t$ for $c = 1$, the expectation value becomes the *MLE* under complete certainty.

In *CertainTrust*, the expectation value E is computed by *fading out* [173] the prior with increasing certainty. The *CertainTrust* certainty function increases as a function of the sample size n and yields the following equation for E:

$$E(t, c, f) := c \cdot t + (1 - c) \cdot f \tag{4}$$

13 This extends, in particular, to scenarios where global stationarity of the data generating process, i.e., the trustee behaviour, cannot be assumed, or the sample is small.

with

$$c = \begin{cases} 0 & \text{if } n = 0 \\ \frac{N \cdot n}{2 \cdot w \cdot (N-n) + N \cdot n} & \text{if } 0 < n < N \\ 1 & \text{if } n \geqslant N \end{cases} \tag{5}$$

for an arbitrary, user established parameter $N \in \{\mathbb{R}^+, +\infty\}$.

In order to maintain a mapping between the *CertainTrust* expectation value, $E(t, c, f)$, and the expectation value of the Beta distribution underlying the Bayesian estimation process, Ries [173] dynamically changes the Bayesian Beta prior. When considering the posterior distribution that results from the application of the *CertainTrust* expectation value computation, the mapping between $E(t, c, f)$ and the expectation value of the Beta posterior is established by making the prior distribution[14] $\mathrm{Beta}(\alpha_0, \beta_0)$ dependent of the sample size n in [173]. That is, it is re-computed for every $n \in \mathbb{R}^+$, resulting in a variable prior for increasing n.

In order to apply this re-computation with a generic certainty estimator[15] $C : (n \in \mathbb{R}^+, \hat{p} \in [0;1]) \mapsto [0;1]$, we provide a generalisation of the formula given by Ries [173] as per the following definition:

Definition 15 (Variable Beta Prior of *CertainTrust* Expectation Value)**.** The Beta prior, $\mathrm{Beta}(\alpha_0, \beta_0)$, for a *CertainTrust* expectation value $E(t, c, f)$ is given for variable $n \in \mathbb{R}^+, t = \hat{p} \in [0;1]$ and a generic certainty estimator $C(n, \hat{p})$ by instantiating α_0, β_0 as

$$\alpha_0 = \begin{cases} f & \text{if } C(n, \hat{p}) = 0 \\ f \cdot (1 - C(n, \hat{p})) \cdot \frac{n}{C(n, \hat{p})} & \text{if } 0 < C(n, \hat{p}) < 1 \\ 0 & \text{if } C(n, \hat{p}) = 1 \end{cases}$$

$$\beta_0 = \begin{cases} (1 - f) & \text{if } C(n, \hat{p}) = 0 \\ (1 - f) \cdot (1 - C(n, \hat{p})) \cdot \frac{n}{C(n, \hat{p})} & \text{if } 0 < C(n, \hat{p}) < 1 \\ 0 & \text{if } C(n, \hat{p}) = 1 \end{cases}$$

For the resulting posterior, $\mathrm{Beta}(\hat{p} \cdot n + \alpha_0, (1 - \hat{p}) \cdot n + \beta_0)$, it holds that

$$E(t, c, f) = E(\mathrm{Beta}(\hat{p} \cdot n + \alpha_0, (1 - \hat{p}) \cdot n + \beta_0)) = \frac{\hat{p} \cdot n + \alpha_0}{n + \alpha_0 + \beta_0}$$

The proof of the equality of the expectation values is given in Appendix C, page 235.

14 Recall that $\mathrm{Posterior} \propto \mathrm{Likelihood} \times \mathrm{Prior}$.

15 Note that we assume that this certainty estimator fulfils the three properties regarding monotonicity and bijection postulated by Wang & Singh [196]; see Property 1 to 3, pp. 59

Assuming a *non-informative prior* under the standard *CertainTrust* instantiation ($f = \frac{1}{2}$, $w = 1$), the prior thus varies between the Uniform prior $\mathrm{Beta}(1,1)$ at $c = 0$ and Haldane's prior $\mathrm{Beta}(0,0)$ at $c = 1$. Thereby, the expectation value increases the rate of convergence between the posterior and frequentist *MLE*, with the latter being the more accurate under the assumption of stationarity [180]. At the same time, the advantages of Bayesian analysis and its interpretation can still be leveraged, in particular the availability of a proper posterior probability distribution of the estimand parameter p.

Under the standard *CertainTrust* model, the parameter c is defined by a function that fades out the a-priori information; in order to avoid confusion with the dispersion based certainty statistics $C_{J;(100\cdot z)\%}$ and $C_{W;(100\cdot z)\%}$, the Jeffreys and Confidence Interval-based Certainty Estimators (Definition 10, p. 59; Definition 12, p. 64), the parameter c in the context of expectation value computation will be referred to as a *fade-out constant* c_e.

The fade-out constant c_e does not directly represent the *statistical reliability* of the point estimate $t = \hat{p}$, but rather the degree of convergence from posterior *MLE* to frequentist *MLE*. This degree of convergence, however, can reasonably be based upon the statistical reliability of $t = \hat{p}$. For this, two distinct cases have to be considered: Fade-out in the limit, that is $N = +\infty$, and fade-out for a fixed $N \in \mathbb{R}^+$, $N \ll +\infty$.

$N = +\infty$: FADE-OUT IN THE LIMIT. Following their construction from interval estimates over the posterior distribution, the Jeffreys and Confidence Interval-based Certainty Estimators (Definition 10, p. 59; Definition 12, p. 64) have a range of $[0;1[$. In the standard *CertainTrust* model, this range maps to setting the parameter N to $+\infty$ [173]. Fading out the (non-informative) prior in the limit – that is, for $\lim_{n\to\infty} \hat{p} = p$ where both $\lim_{n\to\infty} C_{J;(100\cdot z)\%}(x,n) = 1$ and $\lim_{n\to\infty} C_{W;(100\cdot z)\%}(x,n) = 1$ – can be achieved by setting $c_e = C_{J;(100\cdot z)\%}(x,n)$ (depending on the choice of certainty statistic, substitute $c_e = C_{W;(100\cdot z)\%}(x,n)$).

This method provides a fade-out of the influence of the a-priori information in the limit, that for conventional confidence levels[16] is slower than the standard *CertainTrust* certainty measure for $f = \frac{1}{2}$ and $N = +\infty$, given by [173] as:

$$c = \frac{n}{n+2}$$

Additionally, setting $c_e = C_{W;(100\cdot z)\%}(x,n)$ takes the concave shape of the certainty estimators (Definition 10, p. 59; Definition 12, p. 64) into account, fading out the prior considerably quicker for $t = \hat{p} \to 0$ or $t = \hat{p} \to 0$ than for $t = \hat{p} \approx \frac{1}{2}$.

16 Conventional confidence levels, such as $z = 0.95$.

	$c = 0.9$	$c = 0.95$	$c = 0.99$	$c = 0.999$
$1 - z = 0.8$	163	654	16,383	1,638,398
$1 - z = 0.9$	267	1,073	26,894	2,689,597
$1 - z = 0.95$	381	1,533	38,412	3,841,596
$1 - z = 0.99$	659	2656	66,558	6,656,393
$1 - z = 0.995$	782	3,151	78,954	7,896,092
$1 - z = 0.999$	1,072	4,319	108,231	10,824,089

Table 2: Confidence Interval-based Certainty Estimator: Minimum sample size n at $\hat{p} = \frac{1}{2}$ to reach given certainty c with confidence $1 - z$.

$N \ll +\infty$: FADE-OUT FOR A FINITE NUMBER OF REPRESENTATIVE EVIDENCE. A behaviour similar to the standard model with $N \ll +\infty$ can be induced in the two Interval-based Certainty Estimators by determining a minimum certainty threshold that has to be exceeded before setting the certainty parameter to 1. Leveraging the fact that for a fixed sample size, the certainty estimate is minimised at $\hat{p} = \frac{1}{2}$, we can determine the worst case minimum number of experiences necessary to reach a given certainty level. Furthermore, we can alter the rate of convergence by varying the *acceptable residual uncertainty level* $(100 \cdot z)\%$. Table 2 shows the worst case estimates for the Confidence Interval-based Certainty Estimator.

For instance, assume a confidence level of $1 - z = 0.95$ that is conventionally chosen for significance in hypothesis testing. If we set the *desired* certainty to $c = C_{W;5\%}(x, n) = 0.9$, the worst case estimate for N is 381. The interpretation of this is straightforward: to guarantee that 95% of the posterior probability mass are dispersed over no more than an interval length[17] of $1 - c = 1 - 0.9 = 0.1$, we need a sample length of at most 381 (see, Table 2).

The fade-out constant c_e, used for fading out the prior for a finite number of representative evidence, $N \ll +\infty$, can be computed directly (instead of relying on the worst case estimate at $f = \frac{1}{2}$) by leveraging the bijection property of the certainty estimators (Property 3, p. 61). For determining N, the sample size $n > 0$ and the sum of successes x, as well as the desired certainty and confidence level z are known. The fade-out constant c_e is computed in a two-step process:

1. Determine N by computing the inverse interval certainty (see Defintion 13, p. 65) $Z^{-1}(t, c)$, setting $t = \frac{x}{n}$ and c to the *desired* certainty level. This gives the number of representative evidence N.

2. Compute c_e according to Equation 5, p. 70:

17 Recall, the dispersion given by certainty estimate is in the same scale as $\hat{p}$.

$$c_e = \begin{cases} 0 & \text{if } n = 0 \\ \frac{N \cdot n}{2 \cdot w \cdot (N-n) + N \cdot n} & \text{if } 0 < n < N \\ 1 & \text{if } n \geqslant N \end{cases}$$

The modified *CertainTrust* expectation value that fades out the a-priori information for a finite number of representative evidence is given by $E(t, c, f) := c_e \cdot t + (1 - c_e) \cdot f$.

3.1.8 *Extending the Human Trust Interface*

Integrating the interval-based method for certainty estimation into the Human Trust Interface (HTI) [174] is a straightforward process. Figure 2 shows an extended version of the standard HTI.

Recall that the ordinate (horizontal axis) of the HTI records the average rating, i.e., $t = \frac{x}{n}$, of a *CertainTrust* opinion $\omega = (t, c, f)$. The abscissa (vertical axis), in the standard version, gives the certainty, c. The color gradient in the background represents a third dimension, computed as a linear interpolation of RGB color values, that represents the *CertainTrust* expectation value $E(t, c, f) = c \cdot t + (1 - c) \cdot f$. The appearance of the interpolation and distribution of the color values over the graph are dependent on the parameters f and w that determine the assumed Beta prior. For further details, see [173].

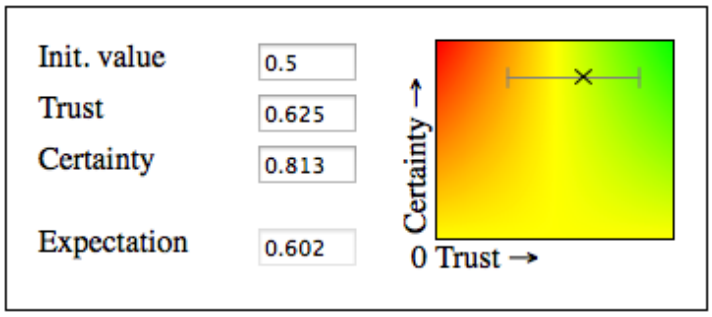


Figure 2: Extended Human Trust Interface (HTI), confidence level: 95 per cent; 5 positive and 3 negative outcomes.

The certainty parameter c of the original version of *CertainTrust* controls the fade-out of the prior knowledge; in the extended version of *CertainTrust* presented above, this parameter is called c_e. In the extension to the HTI proposed here, the prior fade-out is still marked on the vertical axis. However, since the certainty estimate obtained by the novel, dispersion-based interval certainty estimators (Definition 10, p. 59 and Definition 12, p. 64), is on the same scale as the estimate t, the actual dispersion measure of the trust estimate can be presented as a confidence interval in the extended HTI. In Figure 2, this is given as a 95% confidence interval for binary proportions, computed according to Wilson (see, Definition 11, p. 63)[27]. This permits the user to see the potential dispersion, according to the given confidence inter-

val, of the parameter t. By doing so, interface is enriched by another statistical measure in a way that is similar to the standard mathematical practice of displaying confidence intervals on probability density plots.

3.1.9 *Section Summary*

The preceding Section 3.1 has introduced a revised and extended version of Ries' *CertainTrust* model [174]. Particular attention has been paid to a sound derivation of both trust and certainty scores, based upon estimation-theoretic, statistical methods.

The somewhat ad-hoc nature of certainty estimators in the state-of-the-art related work has been addressed and mitigated by introducing two novel, dispersion-based certainty estimators. Derived from credibility and confidence intervals for binomial proportions, the Credibility Interval-based Certainty Estimator (Definition 10) and the Confidence Interval-based Certainty Estimator (Definition 12) leverage established statistical methodology in order to derive a certainty score that reflects the reliability of a given trust estimate. These estimators do not only provide a sound statistical footing for certainty estimation in computational trustworthiness assessment, but also provide a certainty estimate that is on the same scale as the trust estimate. It has been shown above that the two certainty estimators satisfy the necessary monotonicity and bijection relations postulated by Wang & Singh [196].

Additionally, the characteristic of having both certainty and trust estimates at the same scale relies fundamentally on the fact that the dispersion of a point estimate can be measured in terms of its standard deviation. This in turn permits the depiction of uncertainty as an interval estimate in an augmented version of the HTI (Section 3.1.8). Thus, the uncertainty of a trust estimate can be interpreted as the likely dispersion at some given confidence level, resulting in a representation of uncertainty in the form of error-bars around the point estimate. This well-established representation is aimed at increasing the intuitiveness of the graphical representation[18].

Furthermore, the mapping of priors to *CertainTrust* initial trust parameters f and w has been discussed (Section 3.1.6) and the computation of the *CertainTrust* expectation value has been revised in order to incorporate statistically sound certainty estimates (Section 3.1.7). This revision provides *CertainTrust* with statistical sound estimation of uncertainty, from its base estimation model, all the way to the computation of the aggregate expectation value.

18 A closer investigation of whether or not that goal was actually achieved is relegated to future work.

Thus, the prediction and representation model of Ries' *CertainTrust* model [174] has been covered in all of its key aspects and revised where deemed necessary.

In the following Section 3.2, the binomial *CertainTrust* will be extended into the novel *Multinomial CertaintTrust* model.

3.2 MULTINOMIAL CASE

So far, we have considered trust assessment in the binomial case as a *binomial* proportion problem (Section 3.1). As such, it is a special case of an *m-cell multinomial proportion problem* [13, 35, 126]. The binomial case constitutes a 2-cell multinomial proportion problem, in which each experience belongs to one of two exclusive and exhaustive *categories*. In Section 3.1, these categories were called *success* and *failure* and encoded by the numerical values 1 and 0, respectively.

In the binomial case, the evaluation of an interaction – such as rating a product or a service – allows the truster to state whether its expectations of the trustee's behaviour have been met (resulting in a *success*) or not (resulting in a *failure*). However, in many real-world reputation systems, such as the common 5-star rating scale systems, *degrees* of (subjective) expectation fulfilment are encoded in ordered categories. The multinomial model is a commonly used and statistically sound basis for formalising multi-categorical ratings. A number of corresponding models can be found in [113][19].

Extending the model of the 2-cell (binomial) special case to the general m-cell (multinomial) case is straightforward (see, for instance, [12, 13, 35]). As before, suppose a sample $X = \{x_1, x_2, \ldots, x_n\}$ of size $n \in \mathbb{N}$. In the binomial case, each $x_i \in \{0; 1\}$, thereby defining a *category membership* of x_i to category k_1, i.e., *success* (if $x_i = 1$), or k_2, i.e., *failure* (if $x_i = 0$). It was assumed that the sample X was generated by repeated application of a Bernoulli trial with stationary probability of success p, yielding a binomial distribution of the resulting sample, i.e., $X \sim \mathrm{Bin}(n, p)$.

Now, in the m-cell (multinomial) case, assume a generalisation of the repeated Bernoulli trial, in which, instead of just two, the variable x_i has a sample space Ω of $m \in \mathbb{N}, m \geqslant 2$ categories. Furthermore, without loss of generality, assume these m categories to be labeled by a finite range of integers, so that $\Omega = \{1, 2, \ldots, m\}$. Thus, the sample $X \in \{1, 2, \ldots, m\}^n$ is generated by repeated sampling from a *Categorical distribution*[20].

Definition 16 (Categorical Distribution)**.** Let $m > 0$ be the number of outcomes of a random event, each outcome is uniquely labeled as a category by an integer value $1, 2, \ldots, m$. Let p_i be the probability of observing an outcome of category $j \in \{1, 2, \ldots, m\}$. The probability distribution function of the *Categorical distribution* is given by

$$f(x; \mathbf{p}) = \prod_{j=1}^{m} p_j^{[x=j]} \tag{6}$$

19 In the related work, the most prominent example is probably Jøsang's *Subjective Logic* [104] and its companion reputation system, the Dirichlet Reputation System [107], that applies concepts of *Subjective Logic* and extends the Beta Reputation System [108].

20 The Categorical distribution is also referred to as the *Discrete distribution*, e.g., [19].

where $[x = j]$ is the *Iverson Bracket* [71], which evaluates to 1 if $x = j$, 0 if $x \neq j$. Furthermore, $\mathbf{p} = (p_1, p_2, \ldots, p_m)$ and $\sum_{j=1}^{m} p_j = 1$.

Thus, for each trial i, the variable x_i can take on *exactly* 1 out of $m \in \mathbb{N}$ values. There are various alternative representations for such a 1-of-m scheme, for instance, through *o-1 random vectors* [13, 19], where *each trial* i is represented as a *vector* $\mathbf{x}_i = (x_i^1, x_i^2, \ldots, x_i^m)^\intercal$ of length m with the realised category labelled 1 and the $m-1$ unrealised categories labelled 0. This representation is more cumbersome than the one used in Definition 16. It is, however, convenient when considering the marginal distributions of individual categories.

However, irrespective of the representation, the sample resulting from $n \in \mathbb{N}$ repeated categorical trials, $X = \{x_1, x_2, \ldots, x_n\}$, with $x_i \in \{1, 2, \ldots, m\}$, follows a *Multinomial distribution* [19] with stationary parameters $p_1, p_2, \ldots, p_m$. That is, $X \sim \text{Mult}(n, \mathbf{p})$, with $\mathbf{p} = (p_1, p_2, \ldots, p_m)$. The relationship between Categorical distribution and Multinomial distribution is analogous to that of the Bernoulli and Binomial distribution in the binomial case. In fact, while the Categorical distribution is identical to the Bernoulli distribution for $m = 2$ categories, the Multinomial distribution is identical to the Binomial distribution for $m = 2$ categories.

Consequently, the objective of trust assessment in the multinomial case – given a sample $X \in \{1, 2, \ldots, m\}^n$ assumed to be generated by a stationary, categorical random process – is determining the $m \in \mathbb{N}$ parameters $\mathbf{p} = (p_1, p_2, \ldots, p_m)$ of the assumed underlying Multinomial distribution[21].

3.2.1 *Multinomial* CertainTrust *Opinions*

Before trust and certainty estimation in the multinomial case can be addressed, the *CertainTrust* opinion representation has to be suitably extended in order to deal with $m > 2$ categories. The standard, binomial *CertainTrust* opinion representation, $\omega := (t, c, f)$, follows the conventions of representing binomial proportions by only giving the proportion for one of the two possible outcomes. That is, $t = \hat{p}$ is an estimate of the *probability of success* p. The estimate $\hat{q}$ of the complementary *probability of failure* q is omitted.

This is justified because in the binomial case, with only one degree of freedom, the complement is easy to compute, i.e., $\hat{q} = 1 - \hat{p}$. Additionally, we are, by the definition of trust (see [64]), interested primarily in the probability of a positive outcome. That is, in the given setting of trust assessment in the binomial case, we can establish an order of preference over the categories *success* and *failure*, and report only the estimated probability of the preferred outcome, *success*. Because there is no ambiguity over the partitioning of the remaining

21 Note that the Categorical and Multinomial distributions have $m \in \mathbb{N}$ parameters, but $m - 1$ degrees of freedom, because $\sum_{j=1}^{m} p_j = 1$.

Category	Sample $\tilde{X}$ (w/ Length n)						Trust Estimate
	$\vec{x}_1$	$\vec{x}_2$	$\vec{x}_3$	...	$\vec{x}_{n-1}$	$\vec{x}_n$	t
Success	1	0	0	...	1	1	$\hat{p} = \frac{1}{n} \cdot \sum_{i=1}^{n} x_{Succ.,i}$
Failure	0	1	1	...	0	0	$\hat{q} = \frac{1}{n} \cdot \sum_{i=1}^{n} x_{Fail.,i}$
$\sum_{\{Succ.,Fail.\}} \vec{x}_i$	1	1	1	...	1	1	$\hat{p} + \hat{q} = 1$

Table 3: Binomial Trust Assessment in 0-1 Random Vector Form.

probability mass $1 - \hat{p}$ (the entire remaining probability mass is assigned to the single complementary estimate $\hat{q}$), this representation is – arguably – both minimal *and* intuitive.

If we consider trust assessment in the binomial case as a m-cell multinomial proportion estimation problem, with $m = 2$, in a *0-1 random vector* representation, we can explicitly give a proportion for each of the two exclusive and exhaustive categories, *success* and *failure* (Table 3). In order to do so, the binomial sample X is expanded into a $0 - 1$ random vector representation, yielding a modified representation of the sample, $\tilde{X} = \{\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_n\}$. Thus, $\tilde{X}$ is a matrix of dimension $2 \times n$, where each column contains exactly one 1-value and one 0-value. Each row encodes the occurrence of a particular outcome (either *success* or *failure*) over the sample length n. That is, if a *failure* occurs at time i, the corresponding column i of $\tilde{X}$ will be marked: $x_{Success,i} = 0$ and $x_{Failure,i} = 1$.

By applying the point and interval estimation techniques presented in Section 3.1 to each row, we can formulate a multinomial *CertainTrust* opinion representation for the binomial ($m = 2$) special case.

Definition 17 (Multinomial *CertainTrust* Opinion Representation for the Binomial ($m = 2$) Special Case)**.** Let $\omega := (t, c, f)$ be a binomial *CertainTrust* opinion. The corresponding multinomial representation of this opinion, $\omega^{m=2}$, is given by:

$$\omega^{m=2} := ((t_1, c_1, f_1)_1; (t_2, c_2, f_2)_2)$$

where

$$t = t_2 = 1 - t_1,$$
$$c = c_1 = c_2, \text{ and}$$
$$f = f_2 = 1 - f_1.$$

For the binomial special case, $m = 2$, the relation between t_1 and t_2, c_1 and c_2, and f_1 and f_2 follow directly from the shape of Beta distributed posterior. Since the categories, 1 (*failure*) and 2 (*success*), are exclusive and exhaustive[22], the posterior Beta distributions used for

22 Consequently, the proportion estimation problem has $m - 1$ degrees of freedom.

estimating $t_1 = \hat{q} = \frac{1}{n} \cdot \sum_{i=1}^{n} x_{Fail.,i}$ and $t_2 = \hat{p} = \frac{1}{n} \cdot \sum_{i=1}^{n} x_{Succ.,i}$ are mirror images of each other along $p = \frac{1}{2}$.

The multinomial representation of the binomial case can easily be extended to facilitate $m > 2$ categories (Table 4). In this m-cell multinomial case, with $m - 1$ degrees of freedom, m different parameters have to be estimated, yielding the general multinomial opinion representation given in Definition 18.

Definition 18 (Multinomial *CertainTrust* Opinion Representation for the General Case ($m > 2$))**.** Let $\tilde{X}$ be a m-cell multinomial sample with dimension $m \times n$ in $0-1$ random vector form. The corresponding multinomial *CertainTrust* opinion, ω^m, is given by:

$$\omega^m := ((t_1, c_1, f_1)_1; (t_2, c_2, f_2)_2; \ldots; (t_m, c_m, f_m)_m)$$

where

$$\begin{aligned}
&\forall i \in \{1, 2, \ldots, m-1, m\} : t_i = \tfrac{1}{n} \cdot \textstyle\sum_{j=1}^{n} x_{i,j}, \\
&\textstyle\sum_{i=1}^{m} t_i = 1, \\
&\textstyle\sum_{i=1}^{m} f_i = 1, \text{ and} \\
&\forall i \in \{1, 2, \ldots, m-1, m\} : t_i, c_i, f_i \in [0, 1].
\end{aligned}$$

Note, that in Definition 18, the certainty parameters $c_1, c_2, \ldots, c_m$ are not required to be identical and are assigned individually to their corresponding trust estimates $t_1, t_2, \ldots, t_m$. The multinomial *CertainTrust* opinion representation thus permits assigning an independent certainty estimate to each trust estimate; these certainty estimates may be different from each other.

3.2.2 *Multinomial Probability Estimation*

Recalling that $X \sim Mult(n, \mathbf{p})$, with $\mathbf{p} = (p_1, p_2, \ldots, p_m)$, the objective of multinomial trust assessment is the estimation of the $m \in \mathbb{N}$ parameters $\mathbf{p}$ of a Multinomial or Categorical distribution. In order to do so, conjugate prior assumptions of Bayesian statistics can be leveraged. Following the principles of Bayesian statistics, the conjugate prior of the Catagorical and Multinomial distributions is the Dirichlet distribution [126].

This prior can be derived from the Categorical distribution by expressing it as a function of $\mathbf{p} = (p_1, p_2, \ldots, p_m)$, yielding $g(\mathbf{p}) \propto \prod_{i=1}^{m} p_i^{\alpha_i}$ for some parameters $\alpha_1, \alpha_2, \ldots, \alpha_m \in \mathbb{R}^+$. From this, we can derive a probability distribution for the parameters $p_i \in [0; 1]$ (with $\sum_{i=1}^{m} p_i = 1$) by multiplying $g(p)$ with an appropriate normalising constant, so that $\int_0^1 g(p)dp = 1$. The normalising constant is given by:

$$\frac{\Gamma(\sum_{i=1}^{m} \alpha_i)}{\prod_{i=1}^{m} \Gamma(\alpha_i)}$$

Category	Sample $\tilde{X}$ (w/ Length n)						Trust Estimate
	$\vec{x}_1$	$\vec{x}_2$	$\vec{x}_3$	...	$\vec{x}_{n-1}$	$\vec{x}_n$	t
Cat_1	1	0	0	...	1	0	$\hat{p}_1 = \frac{1}{n} \cdot \sum_{i=1}^{n} x_{1,i}$
Cat_2	0	1	0	...	0	0	$\hat{p}_2 = \frac{1}{n} \cdot \sum_{i=1}^{n} x_{2,i}$
$\vdots$				$\ddots$			$\vdots$
Cat_m	0	0	1	...	0	1	$\hat{p}_m = \frac{1}{n} \cdot \sum_{i=1}^{n} x_{m,i}$
$\sum_{\{Cat_1,\ldots,Cat_m\}} \vec{x}_i$	1	1	1	...	1	1	$\sum_{j=1}^{m} \hat{p}_j = 1$

Table 4: Multinomial Trust Assessment in 0-1 Random Vector Form.

Γ denotes the gamma function: $\Gamma(z) = \int_0^\infty t^{z-1} \cdot e^{-t} dt$. The resulting Dirichlet distribution, $\mathrm{Dir}(\mathbf{p}; \alpha_1, \ldots, \alpha_m)$, thus has the following probability density function [35, 126]:

$$f(p_1, p_2, \ldots, p_m; \alpha_1, \alpha_2, \ldots, \alpha_m) = \frac{\Gamma(\sum_{i=1}^{m} \alpha_i)}{\prod_{i=1}^{m} \Gamma(\alpha_i)} \cdot \prod_{i=1}^{m} p_i^{\alpha_i} \tag{7}$$

Supposing a sample $X \sim \mathrm{Mult}(n, \mathbf{p})$ with $m \in \mathbb{N}$ categories, α_i denotes the number of occurrences of the i-th category in X. In other words, when considering sample X in *o-1* random vector form, i.e., $\tilde{X}$ as in Table 4, p. 80, α_i equals the sum over the row representing category i:

$$\alpha_i = \sum_{j=1}^{n} x_{i,j}$$

Estimating the trust scores $t_1, t_2, \ldots, t_m$ is achieved by computing the expectation values of the parameters $p_1, p_2, \ldots, p_m$ of the Dirichlet distributed posterior distribution

$$f(p_1, p_2, \ldots, p_m; \alpha_1 = \sum_{j=1}^{n} (x_{1,j}), \ldots, \alpha_m = \sum_{j=1}^{n} (x_{m,j}))$$

as the following proportions:

$$t_i = \hat{p}_i = \frac{\alpha_i}{n} = \frac{1}{n} \cdot \sum_{j=1}^{n} x_{i,j} \tag{8}$$

By definition, it obviously holds that:

$\sum_{i=1}^{m} \alpha_i = n = |X|$, and
$\sum_{i=1}^{m} \hat{p}_i = 1$

The result is an m-*dimensional multinomial proportion*, computed as posterior maximum likelihood estimates $\hat{\mathbf{p}} = (\frac{\alpha_1}{n}, \frac{\alpha_2}{n}, \ldots, \frac{\alpha_m}{n})$ from a Bayesian Dirichlet posterior. This multinomial proportion represents the trust estimates $\mathbf{t} = (t_1, t_2, \ldots, t_m)$ used to instantiate an m-dimensional *CertainTrust* opinion, $\omega^m = ((t_1, c_1, f_1)_1; \ldots; (t_m, c_m, f_m)_m)$ (Defintion 18, p. 79). The following Section 3.2.3 will extend binomial certainty estimation, discussed in Section 3.1.2, in order to obtain a dispersion-based certainty estimate for each of the $m \in \mathbb{N}$ trust estimates.

3.2.3 *Multinomial Certainty Estimation*

In Defintion 6, p. 53, *certainty* was defined as an estimate for the reliability of the trust estimate $t = \hat{p}$. In the multinomial case, the trust estimate $\mathbf{t} = (t_1, t_2, \ldots, t_m) = (\hat{p}_1, \hat{p}_2, \ldots, \hat{p}_m)$ is an m-dimensional

vector of multinomial proportions, each of which can be assigned its own certainty estimate. The goal of certainty estimation in the m-dimensional multinomial case is to furnish a vector $\mathbf{c} = (c_1, c_2, \ldots, c_m) \in [0;1]^m$ of statistically sound certainty estimates. For this, the certainty estimators presented in Sections 3.1.2 are extended into simultaneous credible/confidence intervals, so as to provide *simultaneous* certainty estimates for multinomial *CertainTrust* opinions.

3.2.4 *Bayesian Interval-Derived Multinomial Certainty*

As outlined in Section 3.2.1, Bayesian estimation of the $m > 2$ parameters $\mathbf{p} = (p_1, p_2, \ldots, p_m)$ of a Multinomial distribution is based on a Dirichlet distributed posterior, $\mathrm{Dir}(\mathbf{p}; \alpha_1, \ldots, \alpha_m)$. The point estimates $(\hat{p}_1, \hat{p}_2, \ldots, \hat{p}_m)$ are computed from this posterior as *marginal proportions*, $\hat{p}_i = \frac{\alpha_i}{n}$. It is of particular interest that we are only considering the margins of the Dirichlet posterior for determining the point estimates, as this allows us to leverage *marginal intervals* in order to compute credible intervals for certainty estimation.

The *marginal distribution* of the i-th parameter, $p_i, 1 \leqslant i \leqslant m$, of $\mathrm{Dir}(\mathbf{p}; \alpha_1, \ldots, \alpha_m)$ is a Beta distribution with:

$$p_i \sim \mathrm{Beta}(\alpha_i, (\sum_{j=1}^{m} \alpha_j) - \alpha_i)$$

In order to compute a Bayesian interval-derived certainty estimate for the $m \in \mathbb{N}$ individual marginal proportions $\hat{p}_1, \hat{p}_2, \ldots, \hat{p}_m$, the Credibility Interval-based Certainty Estimator (Definition 10, p. 59) is applied to each marginal proportion. The result is a certainty estimate c_i for the trust estimate $t_i = \hat{p}_i$:

$$c_i = C_{J;(100\cdot\tilde{z})\%}(x = \alpha_i, n = \sum_{j=1}^{m} \alpha_j) := 1 - (U_J(x) - L_J(x))$$

For the individual certainty estimates c_i, the properties regarding monotonicity and bijection (Properties 1, p. 59, to 3, p. 61) still hold. Note also, that the residual acceptable uncertainty – defining the confidence level used to compute the Jeffreys interval certainty estimate – denoted as $\tilde{z}$ instead of z. This is owed to the fact that, in order to correct for multiple testing[23] when computing $m \in \mathbb{N}$ simultaneous certainty estimates, the individual confidence levels, i.e., $1 - \tilde{z}$, have to be adjusted appropriately to guarantee a confidence of $1 - z$ for the entire m-dimensional simultaneous certainty vector $c = (c_1, c_2, \ldots, c_m)$.

In formal terms, $\tilde{z}$ should be chosen in such a manner that the simultaneous coverage probability is at least $1 - z$:

$$P(p_i \in [L_i; U_i] : \forall i = 1, \ldots, m) \geqslant 1 - z$$

23 For an overview on multiple testing, see, for instance, [90].

L_i and U_i represent the lower and upper bounds of the credible interval for the i-th parameter, p_i. Both the *Bonferroni* and the *Šidák* adjustments [156] provide conservative[24] solutions for determining $\tilde{z}$.

The *Bonferroni adjustment* maintains the simultaneous coverage probability by constructing the $m > 2$ different marginal intervals with an adjusted confidence level of $\tilde{z} = \frac{z}{m}$. However, Bonferroni adjustment can be excessively conservative [156] for large $m \in \mathbb{N}$.

The *Šidák adjustment* is slightly more powerful than the Bonferroni adjustment, but requires *mutual independence* [13] of the estimates. This requirement is met under the assumption of a stationary Categorical process behind the generation of sample X for the estimates $\hat{p}_1, \ldots, \hat{p}_m$. The Šidák adjustment is achieved by setting the adjusted confidence level $\tilde{z} = 1 - (1 - z)^{\frac{1}{m}}$.

Since the number of categories encountered in trust assessment tends to be relatively small[25] and because the Bonferroni method is more general by not requiring independence[26] of the estimates, the use of the simpler Bonferroni adjustment is advocated in the following. Consequently, the adjusted Simultaneous Credibility Interval-based Certainty Estimator for multinomial proportions with $m > 2$ categories can be defined as a vector of simultaneous certainty estimates.

Definition 19 (Simultaneous Credibility Interval-based Certainty Estimator for Multinomial Proportions)**.** The *Simultaneous Credibility Interval-based Certainty Estimator* for an m-dimensional ($m > 2$) multinomial trust estimate $\mathbf{t} = (t_1, t_2, \ldots, t_m) = (\frac{\alpha_1}{n}, \frac{\alpha_2}{n}, \ldots, \frac{\alpha_m}{n})$ and an *acceptable residual uncertainty level* (confidence level) $(100 \cdot z)\%$, is defined as

$$C^m_{J;(100\cdot z)\%}(\alpha_1, \alpha_2, \ldots, \alpha_m; n) = (c_1, c_2, \ldots, c_m)$$

where

$$c_i = C_{J;(100\cdot \tilde{z})\%}(\alpha_i, n) := 1 - (U_J(\alpha_i) - L_J(\alpha_i))$$

with

$$\tilde{z} = \frac{z}{m}$$

Because the individual certainty estimates $c_1, c_2, \ldots, c_m$ are Bonferroni adjusted marginal Credibility Interval-based Certainty Estimates, adhering to Definition 10, p. 59, each marginal certainty estimate accords with the properties postulated by Wang & Singh [196] regarding monotonicity for fixed $\hat{p}$ (Property 1) and for fixed n (Property 2).

24 Conservative in the sense of guaranteeing that $P(p_i \in [L_i; U_i] : \forall i = 1, \ldots, m) > 1 - z$.

25 For instance, $m = 5$ (*Amazon Star Ratings*), $m = 10$ (*IMDB movie ratings*).

26 The independence assumption is a property of the theoretical model, but might be violated in real world applications.

For the same reason, the algorithm applied to find an inverse function $(x, n) = Z^{-1}(t, c)$ from opinion to evidence space (Algorithm 1, p. 62) [197] is also applicable in the multinomial case by substituting $\tilde{z}$ for z (Algorithm 2).

Data: Trust estimate $\mathbf{t} = (t_1, t_2, \ldots, t_m)$, $t_i = \frac{\alpha_i}{n}$, certainty estimate $\mathbf{c} = (c_1, c_2, \ldots, c_m)$, $c_i = C_{J;(100\cdot\tilde{z})\%}(\alpha_i, n)$, accpetable residual uncertainty (confidence) z
Result: Number of occurrences per category $\alpha_1, \alpha_2, \ldots, \alpha_m$, sample size n

// Initialize parameters
$\mathbf{t} = (t_1, t_2, \ldots, t_m)$;
$\mathbf{c} = (c_1, c_2, \ldots, c_m)$;
$\tilde{z} = \frac{z}{m}$;
$n = 0$;
$n_1 = 0$;
$n_2 = n_{max}$;
// Select arbitrary t_i *for determining* n*, e.g.,* t_1
// Approximate n *to specified precision* ϵ*, as in Alg. 1*
while $n_2 - n_1 \geqslant \epsilon$ **do**
  $n = \frac{n_1+n_2}{2}$;
  if $C_{J;(100\cdot\tilde{z})\%}(t_1 \cdot n, n) < c_1$ **then**
    $n_1 = n$
  end
  else
    $n_2 = n$
  end
end
return $n, \alpha_1 = t_1 \cdot n, \alpha_2 = t_2 \cdot n, \ldots, \alpha_m = t_m \cdot n$

Algorithm 2: Calculation of $(n, \alpha_1, \alpha_2, \ldots, \alpha_m) = Z^{-1}(\mathbf{t}, \mathbf{c})$ (see also [197] and Algorithm 1, p. 62)

Algorithm 2 is of the same complexity as the equivalent solution for the binomial case (Agorithm 1) because the core part, the approximation of sample size n, has to be executed only once. However, just as in the binomial case, the Simultaneous Credibility Interval-based Certainty Estimator cannot be represented in closed form. A closed form approximation, based on an extension of the Wilson confidence intervals to the mutlinomial case, is formulated in the next section, paralleling the construction of the Wilson Interval Certainity Estimator (Definition 12, p. 64).

3.2.5 *Confidence Interval-Derived Multinomial Certainty*

The Wilson confidence interval [204] for binomial proportions is a modified normal approximation of the distribution of the error of $\hat{p}$,

that can be derived from Rao's score test [169]. It is defined as (see [27] and Definition 11, p. 63):

$$\mathrm{CI}_W = \frac{x + \frac{\kappa^2}{2}}{n + \kappa^2} \pm \frac{\kappa \cdot \sqrt{n}}{n + \kappa^2} \cdot \sqrt{\hat{p} \cdot (1 - \hat{p}) + \frac{\kappa^2}{4 \cdot n}}$$

where κ is the $100 \cdot (1 - \frac{z}{2})$ percentile of the standard normal distribtution, i.e., $\kappa = \Phi^{-1}(1 - \frac{z}{2})$.

The standard normal distribution is a special case of the chi-squared distribution. The chi-squared distribution with m degrees of freedom describes the distribution of the sum of $m \in \mathbb{N}$ independently distributed standard normal random variables (see, for instance, [126, 148]). Consequently, the standard normal distribution can be expressed as a chi-squared distribution with one degree of freedom. The $1 - \frac{z}{2}$ percentile can be determined equivalently from both distributions[27], leading to $\kappa = \Phi^{-1}(1 - \frac{z}{2}) = \chi^2(\frac{z}{2}, 1)$.

Quesenberry & Hurst [165] leverage this relationship between the Wilson interval and Pearson's chi-squared statistic [161] for constructing simultaneous intervals for multinomial proportions. For doing so, they propose using $\kappa = \chi^2(\frac{z}{2}, m-1)$, for a multinomial proportion estimation problem with m categories.

Goodman [70] extends [165] by invoking the Bonferroni argument [137]. Goodman simultaneous confidence intervals are constructed by setting $\kappa = \chi^2(\frac{z}{2 \cdot m}, 1)$, which provides shorter, yet still conservative, intervals at a confidence that is closer to the nominal confidence levels [70]. This makes the Goodman intervals preferable over the Quesenberry & Hurst intervals for the multinomial case ($m > 2$) [137].

From the Goodman simultaneous confidence interval, the *Simultaneous Confidence Interval-based Certainty Estimator for Multinomial Proportions*, extending the *Confidence Interval-based Certainty Estimator* (Definition 12, p. 64), is derived in the following Definition 20:

Definition 20 (Simultaneous Confidence Interval-based Certainty Estimator for Multinomial Proportions)**.** The *Simultaneous Confidence Interval-based Certainty Estimator* for an m-dimensional ($m > 2$) multinomial trust estimate $\mathbf{t} = (t_1, t_2, \ldots, t_m) = (\frac{\alpha_1}{n}, \frac{\alpha_2}{n}, \ldots, \frac{\alpha_m}{n})$ and an *acceptable residual uncertainty level* (confidence level) $(100 \cdot z)\%$, is defined as

$$C^m_{G;(100 \cdot z)\%}(\alpha_1, \alpha_2, \ldots, \alpha_m; n) = (c_1, c_2, \ldots, c_m)$$

where

$$c_i := 1 - (U_G(\alpha_i) - L_G(\alpha_i))$$

with

$$U_G(\alpha_i) = \frac{\alpha_i + \frac{\kappa^2}{2}}{n + \kappa^2} + \frac{\kappa \cdot \sqrt{n}}{n + \kappa^2} \cdot \sqrt{\frac{\alpha_i}{n} \cdot (1 - \frac{\alpha_i}{n}) + \frac{\kappa^2}{4 \cdot n}},$$
$$L_G(\alpha_i) = \frac{\alpha_i + \frac{\kappa^2}{2}}{n + \kappa^2} - \frac{\kappa \cdot \sqrt{n}}{n + \kappa^2} \cdot \sqrt{\frac{\alpha_i}{n} \cdot (1 - \frac{\alpha_i}{n}) + \frac{\kappa^2}{4 \cdot n}}$$

27 That is, in tabulated form, it can be looked up in either the normal or the chi-squared percentile tables.

and

$$\kappa = \chi^2(\frac{z}{2 \cdot m}, 1).$$

From the construction of the *Simultaneous Confidence Interval-based Certainty Estimator for Multinomial Proportions* as an extension of the *Confidence Interval-based Certainty Estimator*, it is evident that Properties 1, p. 59, to 3, p. 61 still hold for the individual certainty estimates $c_1, \ldots, c_m$. The inverse relation $Z^{-1}(\mathbf{t}, \mathbf{c})$, $\mathbf{t} = (t_1, t_2, \ldots, t_m)$, $\mathbf{c} = (c_1, c_2, \ldots, c_m)$ can be given in closed form for the *Simultaneous Confidence Interval-based Certainty Estimator*, analogously to the inverse in the binomial case presented in Definition 13, p. 65.

Definition 21 (Inverse Simultaneous Confidence Interval-based Certainty)**.** Let a multinomial *CertainTrust* opinion ω^m of dimension $m > 2$ (Definition 18), $\omega^m := ((t_1, c_1, f_1)_1; \ldots; (t_m, c_m, f_m)_m)$, be given. Furthermore, let $(100 \cdot z)\%$, the acceptable residual uncertainty level under which the certainty estimates $\mathbf{c} = (c_1, \ldots, c_m)$ were computed, be known and correspondingly, let κ be the chi-squared value obtained from $\chi^2(\frac{z}{2 \cdot m}, 1)$. The relation $Z^{-1}(\mathbf{t}, \mathbf{c}) = (n, \alpha_1, \alpha_2, \ldots, \alpha_m)$ is given by:

$$n = \frac{-\kappa^2 \cdot \left(2u^2 - 4 \cdot t_1 + 4 \cdot t_1^2\right)}{2u^2} + \frac{\sqrt{4u^2 \cdot \kappa^4 \cdot (1 - u^2) + \kappa^4 \cdot \left(2u^2 - 4 \cdot t_1 + 4 \cdot t_1^2\right)^2}}{2u^2}$$

$$\alpha_i = t_i \cdot n$$

with $u = 1 - c_1$ (i.e., the length of the Goodman interval for proportion $t_1 = \frac{\alpha_1}{n}$).

The *Simultaneous Confidence Interval-based Certainty Estimator* provides closed form computation of a multinomial certainty estimate, that is only marginally more complex than the computation of its binomial analogue, the *Confidence Interval-based Certainty Estimator*. The same holds for the inverse relation given in Definition 21.

3.2.6 *Initial Trust Value*

In Section 3.1.6, the instantiation of a Bayesian Beta prior distribution from the *CertainTrust* parameters f and w was discussed for the Binomial case. For the Multinomial case, an analogous instantiation from the *Multinomial CertainTrust* parameters $f_1, \ldots, f_m$ and w to the conjugate prior distribution of the Multinomial/Categorical distributions, the Dirichlet prior, must also be established.

Definition 22 (Initial instantiation of Dirichlet prior with Multinomial *CertainTrust* parameters)**.** Let α_i^0 denote the subjective, non-frequentist

prior information corresponding to category i, expressed as pseudo-counts. The initial Dirichlet prior, $f(p_1, p_2, \ldots, p_m; \alpha_1^0, \alpha_2^0, \ldots, \alpha_m^0)$, for an m-dimensional ($m \geqslant 2$), multinomial trust estimation problem is instantiated from *Multinomial CertainTrust* parameters $f_1, f_2, \ldots, f_m$ and w in the following manner:

$$\begin{aligned} \alpha_i^0 &= m \cdot w \cdot f_i, \\ f(p_1, p_2, \ldots, p_m; \alpha_1^0, \alpha_2^0, \ldots, \alpha_m^0) &= \frac{\Gamma(\sum_{i=1}^{m} \alpha_i^0)}{\prod_{i=1}^{m} \Gamma(\alpha_i^0)} \cdot \prod_{i=1}^{m} p_i^{\alpha_i^0} \end{aligned}$$

The *Multinomial CertainTrust* parameters $f_1, \ldots, f_m$ and w determine the shape of the prior Dirichlet distribution. In the multinomial case, $f_i \in [0; 1]$, $\sum_{i=1}^{m} f_i = 1$ encodes a multinomial proportion, while $m \cdot w \in \mathbb{R}^+$ represents a *pseudo count* of subjective experiences that is partitioned according to $f_1, \ldots, f_m$. Just as in the binomial case, their concrete choice determines whether or not the resulting prior distribution is an *informative* or a *non-informative* prior. Informative priors, as has been discussed for the binomial case already, encode subjective *a priori* knowledge. In the multinomial case of trust estimation, informative initial priors are characterised by $\neg(f_i = \frac{1}{m} : \forall f_i \in \{f_1, \ldots, f_m\})$ or $w > 1$.

The reference priors most commonly encountered in multinomial proportion estimation problems – the Uniform, Haldane's, Jeffreys and Perk's priors (see, for instance, [40]) – are instantiated from $f_1, \ldots, f_m$ and w as described in the following.

- *Uniform Prior*: $\mathrm{Dir}(\alpha_1 = \alpha_2 = \ldots = \alpha_m = 1) \leftrightarrow f_1 = f_2 = \ldots = f_m = \frac{1}{m}, w = 1$.
- *Haldane's Prior*: $\mathrm{Dir}(\alpha_1 = \alpha_2 = \ldots = \alpha_m = 0) \leftrightarrow f_1 = f_2 = \ldots = f_m = \frac{1}{m}, w = 0$.
- *Jeffreys Prior*: $\mathrm{Dir}(\alpha_1 = \alpha_2 = \ldots = \alpha_m = \frac{1}{2}) \leftrightarrow f_1 = f_2 = \ldots = f_m = \frac{1}{m}, w = \frac{1}{2}$.
- *Perk's Prior*[28]: $\mathrm{Dir}(\alpha_1 = \alpha_2 = \ldots = \alpha_m = \frac{1}{m}) \leftrightarrow f_1 = f_2 = \ldots = f_m = \frac{1}{m}, w = \frac{1}{m}$.

Jøsang & Haller [107] propose the use of the Uniform prior for their *Dirichlet Reputation System*, justifying their choice by the uniformity criterion over the parameter space. That is, the uniform prior yields a 'flat' distribution that assigns the same probability to all values of the parameter space. This is a reasonable line of argumentation. Given that the uniform prior is a very popular choice in Bayesian trust models, its use as a reference prior in *Multinomial CertainTrust* is suggested.

However, it should be noted that the choice of reference prior is a matter of convention [117], and the non-informativity of a prior

28 In the binomial case, $m = 2$, Perk's prior obviously coincides with Jeffreys prior.

can be measured in various ways – such as Fisher information [99] or entropy [98]. Jeffreys, Perk's and Haldane's priors are more *MLE*-favouring[29], compared to the Uniform prior. For large sample sizes $n \in \mathbb{N}$, the unbiased frequentist *MLE* is the most accurate estimator. While the impact of the prior automatically decreases with increasing sample size n, the bias induced by a prior with $w \neq 0$ can be removed entirely by fading out the prior. For this, the expectation value computation with variable prior presented in Section 3.1.7, p. 69 is applied to the multinomial case.

3.2.7 *Adjusted Expectation Value Computation*

For an m-dimensional multinomial trust estimation problem, the expectation value of the corresponding m-dimensional *Multinomial CertainTrust* opinion, $E(\omega^m) = E((t_1, c_1, f_1), \ldots, (t_m, c_m, f_m))$, becomes an m-dimensional vector of expectations values:

$$E(\omega^m) = E((t_1, c_1, f_1)_1; \ldots; (t_m, c_m, f_m)_m) = (E_1(t_1, c_1, f_1), \ldots, E_m(t_m, c_m, f_m))$$

Each of the constituents of the *Multinomial CertainTrust* opinion, i.e., the individual triples $(t_i, c_i, f_i), i \in \{1, 2, \ldots, m\}$, represents a marginal[30] proportion, $t_i = \frac{\alpha_i}{n}$, and its concordant certainty estimate, c_i. Therefore, the adjusted expectation value computation from Section 3.1.7 can be applied to each constituent triple. However, in the multinomial case with $m > 2$ categories and an arbitrary certainty estimator, it cannot be guaranteed that $\sum_{i=1}^{m} E_i(t_i, c_i, f_i) = 1$ Thus, the computation of the individual expectation values is given by applying a normalisation:

$$\forall i \in \{1, 2, \ldots, m\} : E_i(t_i, c_i, f_i) := \frac{c_i \cdot t_i + (1 - c_i) \cdot f_i}{\sum_{j=1}^{m} (c_j \cdot t_j + (1 - c_j) \cdot f_j)}$$

Again, let α_i^0 denote the subjective, non-frequentist component corresponding to $\alpha_i, i \in \{1, 2, \ldots, m\}$, expressed as a pseudo-count. The marginal Beta distributions of a Bayesian Dirichlet posterior are given by:

$$\mathrm{Beta}(\alpha_i + \alpha_i^0, (\sum_{j=1}^{m} (\alpha_j + \alpha_j^0)) - (\alpha_i + \alpha_i^0)) \tag{9}$$

The generic certainty estimator $C(n, t_i)$ may be dependent on the sample length $n \in \mathbb{N}$ *and* the marginal proportion $t_i = \frac{\alpha_i}{n}$. Such certainty estimators – for instance, the *Confidence Interval-based Certainty Estimator* (Definition 12, p. 64) and its multinomial analogue, the *Simultaneous Confidence Interval-based Certainty Estimator* (Definition 20, p. 85) – can yield concave certainty functions. Consequently,

29 In ascending order; that is: Jeffreys prior is the most biased, while Haldane's prior will result in the frequentist *MLE* itself.

30 Recall that the in the margins the conjugate Dirichlet prior simplifies to a Beta prior.

the individual certainty estimates $c_1, \ldots, c_m$ contain information on the marginal proportions $t_1, \ldots, t_m$, because t_i is a component of the variance of the marginal Beta distribution (Equation 3, p. 56). While in the binomial case, the certainty estimates for category *success* and category *failure* are identical – the Beta distribution has only one degree of freedom – the individual simultaneous certainty estimates in a multinomial setting are generally not identical.

An adaptation of the method for computing the variable prior presented in Definition 15, p. 70, would require solving an m-dimensional system of equations in order to determine the interdependent variable prior components $\alpha_1^0, \ldots, \alpha_m^0$:

$$\alpha_i^0 = \frac{E_i(t_i, c_i, f_i) \cdot (n + \sum_{j=1}^{i-1} \alpha_j^0 + \sum_{j=i+1}^{i-m} \alpha_j^0)}{1 - E_i(t_i, c_i, f_i)} \tag{10}$$

under the constraints of $\alpha_i^0 \geqslant 0$, $\sum_{i=1}^{m} E_i(t_i, c_i, f_i) = 1$.

Depending on the choice of certainty estimator to compute the certainty estimate c_i, computing the $m \in \mathbb{N}$ different α_i is infeasible. Additionally, the variable prior that Equation 10 produces does not generally maintain non-informativity, even if $f_1 = f_2 = \ldots = f_m = \frac{1}{m}$. In other words, for the proportions between the *Multinomial CertainTrust* parameters $\mathbf{f} = (f_1, f_2, \ldots, f_m)$ it does *not* generally hold that:

$$\frac{f_i}{f_j} = \frac{\alpha_i + \alpha_i^0}{\sum_{k=1}^{m}(\alpha_k + \alpha_k^0)} \cdot \frac{\sum_{k=1}^{m}(\alpha_k + \alpha_k^0)}{\alpha_j + \alpha_j^0} = \frac{\alpha_i + \alpha_i^0}{\alpha_j + \alpha_j^0}$$

In order for the prior fade-out to occur uniformly over the expectation values for all $m \in \mathbb{N}$ categories, $(E_1(t_1, c_1, f_1), \ldots, E_m(t_m, c_m, f_m))$, a norm can be applied to the m different estimates. A minimum norm over the certainty vector $\mathbf{c} = (c_1, c_2, \ldots, c_m)$ yields a conservative fade-out parameter $c_e = \min(c_1, c_2, \ldots, c_m)$.

However, the estimate that is arguably of the most interest is the one for the category that is the most likely to occur. That is, we will set c_e in such a manner, so that it corresponds to the certainty estimate c_i of the category i, from $m \in \mathbb{N}$ categorical alternatives, which has the highest trust estimate t_i.

$$c_e = c_i, i \text{ so that } t_i = \max(t_1, t_2, \ldots, t_m) \tag{11}$$

Fading out the prior uniformly across all $m \in \mathbb{N}$ categories by, choosing c_e as described, results in the following Definition 23:

Definition 23 (Variable Dirichlet Prior of *Multinomial CertainTrust* Expectation Value)**.** The Dirichlet prior, $\mathrm{Dir}(\alpha_1^0, \alpha_2^0, \ldots, \alpha_m^0)$, for a m-dimensional *Multinomial CertainTrust* expectation value $E(\omega^m)$ is given for variable $n \in \mathbb{R}^+, \mathbf{t} = (t_1, t_2, \ldots, t_m), t_i \in [0; 1], \sum_{i=1}^{m} t_i = 1, \mathbf{f} = (f_1, f_2, \ldots, f_m), f_i \in [0; 1], \sum_{i=1}^{m} f_i = 1$ and a generic certainty estimator $C(n, t_i) = c_i$, $\mathbf{c} = (c_1, c_2, \ldots, c_m)$ by instantiating the fade-out constant c_e as in Equation 11 and $\alpha_i^0, i \in \{0, 1, \ldots, m\}$ as

$$\alpha_i^0 = \begin{cases} f_i & \text{if } c_e = 0 \\ f_i \cdot (1 - c_e) \cdot \frac{n}{c_e} & \text{if } 0 < c_e < 1 \\ 0 & \text{if } c_e = 1 \end{cases}$$

The proof for $E_i(t_i, c_i, f_i) = c_e \cdot t_i + (1 - c_e) \cdot f_i = \frac{\alpha_i + \alpha_i^0}{\sum_{j=1}^{m}(\alpha_j + \alpha_j^m)}$ follows directly from Equation 9, p. 88, and the proof for the binomial case, Appendix C, p. 235.

The method for fading-out the a-priori information in the multinomial case that was presented so far only considered a fade-out for N (the minimum number of representative evidence) approaching infinity. However, for a fade-out with fixed $N \ll +\infty$, the method introduced in Section 3.1.7, p. 72 can be suitably adapted.

1. Determine N by choosing c at the *desired* certainty level, setting $t_{max} = max(t_1, t_2, \ldots, t_m)$ and computing the inverse of the Goodman Certainty Estimator (Definition 21, p. 86):

$$N = \frac{-\kappa^2 \cdot \left(2u^2 - 4 \cdot t_{max} + 4 \cdot t_1^2\right) + S}{2u^2}$$

with $u = 1 - c$ (i.e., the length of the Goodman interval for proportion $t_{max} = max(\mathbf{t})$) and

$$S = \sqrt{4u^2 \cdot \kappa^4 \cdot (1 - u^2) + \kappa^4 \cdot (2u^2 - 4 \cdot t_{max} + 4 \cdot t_{max}^2)^2}$$

2. Compute c_e according to Equation 5, p. 70:

$$c_e = \begin{cases} 0 & \text{if } n = 0 \\ \frac{N \cdot n}{2 \cdot w \cdot (N - n) + N \cdot n} & \text{if } 0 < n < N \\ 1 & \text{if } n \geqslant N \end{cases}$$

Consequently, the corresponding, m-dimensional expectation vector

$$E(\omega^m) = E((t_1, c_1, f_1)_1; \ldots; (t_m, c_m, f_m)_m) = (E_1(t_1, c_1, f_1), \ldots, E_m(t_m, c_m, f_m))$$

can be computed from $E_i(t_i, c_i, f_i) = c_e \cdot t_i + (1 - c_e) \cdot f_i$ with a fixed minimum number of representative N.

3.2.8 *Representing* Multinomial CertainTrust *Opinions in the HTI*

Comparing two *Multinomial CertainTrust* opinions and deciding which one is "better" in general requires a multidimensional optimality criterion, such as Pareto optimality – if no further assumptions are introduced. However, in the application area considered here, i.e., multinomial reputation and trust systems, it is generally assumed that the feedback categories, that are compounded into the multinomial evidence in sample vector $X \sim \mathrm{Mult}(n, \mathbf{p})$, are forming a strictly ordered set.

For instance, the ubiquitous 5-star reputation system[31], relies on 5 strictly ordered feedback categories, where a 5-star rating is *better* than a 4-star rating, which is *better* than a 3-star rating, and so on. The exact definition *better* is of no relevance here; it might be the presence or absence of a particular feature, as in hotel association stars, or a subjective degree of satisfaction experienced by a customer, as in product rating sites.

In the preceding sections, the $m \in \mathbb{N}$ categories of a multinomial trust estimation problem have been index by positive integers $1, 2, \ldots, m$. In the following, it will be supposed that the order relation *greater* $- > -$ on the set of integers $\{1, 2, \ldots, m\}$, which imposes a strict order over the elements of $\mathbb{N}$, represents the order of the "goodness" of the categories. That is, category c_i is *better* than category c_j, if $i > j$ for $i, j \in \mathbb{N}$.

Additionally, the aggregation property for Dirichlet distributed random variables (see, for instance, [66, 110]) can be leveraged. In fact, the aggregation property has already been used to compute the marginal Beta distributions of the Dirichlet posterior in Equation 9, p. 88.

Let $\mathbf{t} = (t_1, t_2, \ldots, t_m) \sim \mathrm{Dir}(\alpha_1, \alpha_2, \ldots, \alpha_m)$. Then, if the trust estimates with indices i and j, t_i and t_j, are replaced by their sum $t_i + t_j$, it holds that $\mathbf{t}' = (t_1, \ldots, t_i + t_j, \ldots, t_m) \sim \mathrm{Dir}(\alpha_1, \ldots, \alpha_i + \alpha_j, \ldots, \alpha_m)$. This, in fact, holds for any non-trivial partitioning and reordering of $\mathbf{t}$. The proof[32]is given in Appendix C, p. 235.

Combining the assumption of a strictly ordered set of $m \in \mathbf{N}$ mutually exclusive categories and the aggregation property of the Dirichlet distribution, it is now straightforward to extend the binomial representation in the HTI so that it will provide support for *Multinomial CertainTrust* Opinions. Jøsang & Haller [107] propose using a histogram to depict the relative or absolute frequency of occurrences of different categories. This is a flexible and conventionally used approach to represent categorical data[33].

31 Found, for example, on Amazon.com product pages.

32 This proof can also be found in various variants in standard textbooks and papers – for instance, [60, 110].

33 It is, for instance, used in Amazon.com product ratings to illustrate the distribution of multi categorical ratings.

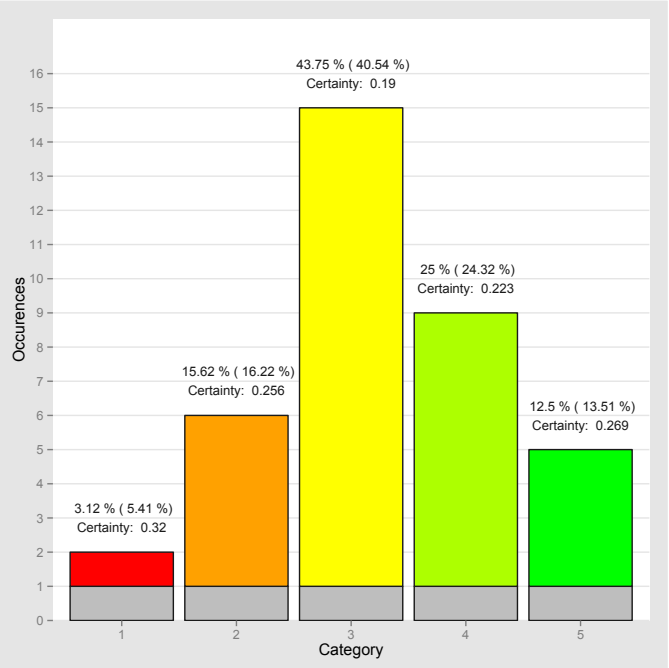


Figure 3: Multinomial opinion representation (5 Categories), with uniform prior.

Figure 3 shows an exemplary multinomial rating of dimension $m = 5$, represented as a histogram over the categories. Note, that these categories are strictly ordered from category 1 (*worst*, red) to category 5 (*best*, bright green). Furthermore, the grey bases of the categorical columns in the chart represent the prior, in this case a Uniform prior. The frequentist MLE is given in per cent, as is the posterior MLE with the Uniform prior (in parentheses). Additionally, the Simultaneous Confidence Interval-based Certainty Estimate for each of the marginal proportions at the 95% confidence level is shown for each of the 5 marginal proportions. A way of combining the histogram representation with the *Human Trust Interface* (HTI) – in order to represent multinomial opinions in the HTI – is shown in Figure 4. Here, both the order over the categories, form *worst* to *best*, and the partitioning property of the Dirichlet distribution are explicitly leveraged.

Assume, for now, that the $m = 5$ different, mutually exclusive categories represent subjective degrees of satisfaction. Of these categories, each of which has a corresponding probability of occurring estimated by $t_i = \frac{\alpha_i}{\sum_{j=1}^{m} \alpha_j}$, with $i \in \{1, 2, \ldots, m\}$ and α_i being the counts of occurrences in category i, as before. Using the partitioning property of the Dirichlet distribution, different relations can now be defined, with regard to an expected degree of satisfaction. This results in binary partitionings in the following manner:

- *Probability that Outcome* $\geqslant$ *Category* i: $\tilde{\alpha} = \sum_{j=i}^{m} \alpha_j$; $\tilde{\beta} = \sum_{j=1}^{i-1} \alpha_j$; $p(x; \tilde{\alpha}, \tilde{\beta}) = \mathrm{Dir}(\tilde{\alpha}, \tilde{\beta}) = \mathrm{Beta}(\tilde{\alpha}, \tilde{\beta})$
- *Probability that Outcome* $=$ *Category* i: $\tilde{\alpha} = \alpha_i$; $\tilde{\beta} = \sum_{j=1;j\neq i}^{m} \alpha_j$; $p(x; \alpha_i, \tilde{\beta}) = \mathrm{Dir}(\alpha_i, \tilde{\beta}) = \mathrm{Beta}(\alpha_i, \tilde{\beta})$ – this is the marginal distribution of α_i
- *Probability that Outcome* $\leqslant$ *Category* i: $\tilde{\alpha} = \sum_{j=1}^{i} \alpha_j$; $\tilde{\beta} = \sum_{j=i+1}^{m} \alpha_j$; $p(x; \tilde{\alpha}, \tilde{\beta}) = \mathrm{Dir}(\tilde{\alpha}, \tilde{\beta}) = \mathrm{Beta}(\tilde{\alpha}, \tilde{\beta})$

Of course, other relations are also easily defined in a similar manner, such as $>, <$ or membership of the outcome to an arbitrary binary partitioning of $(1, 2, \ldots, m)$. Any binary partitioning can be represented as a Beta distribution, according to the partitioning property. This is, in turn, compatible with the representation of binomial opinions in *CertainTrust*. In fact, the variables $\tilde{\alpha}$ and $\tilde{\beta}$ are the sums of Dirichlet distributed random variables and can serve as input parameters for the trust value representation underlying the HTI.

Figure 4: Combining the HTI and histogram opinion representations.

By defining the relation to be represented (e.g., $\geqslant$ or $\leqslant$) and selecting the desired category at which the binary partitioning is to occur, a new binomial CertainTrust opinion is generated. This can be achieved by selecting the desired category in the histogram, for instance by clicking on it (Figure 4, blue arrow) and displaying the resulting binomial opinion in the HTI.

The new binomial CertainTrust opinion, $\omega = (t, c, f)$, is defined by $t = \frac{\tilde{\alpha}}{\tilde{\alpha}+\tilde{\beta}}$, $c = C(\tilde{\alpha} + \tilde{\beta}, t)$ for an arbitrary certainty estimator $C(n, p)$ (Definition 6,p. 53). The value of f depends on the kind of prior distribution that is assumed to underly the Dirichlet posterior. In case of a non-informative prior, a new non-informative Beta prior should be chosen, resulting in $f = 0.5$ and $w \leqslant 1$. In case the prior is informative, that is, it contains actual information on the assumed multinomial distribution, f can for instance be computed in the following manner.

Let α_i^0 be the prior component corresponding to category count α_i. Accordingly, let $\tilde{\alpha}^0$ and $\tilde{\beta}^0$ be the sums of those α_i^0 that belong to $\tilde{\alpha}$ and $\tilde{\beta}$, respectively. In this case,

$$f = \frac{\tilde{\alpha}^0}{\tilde{\alpha}^0 + \tilde{\beta}^0}$$

and

$$w = \tilde{\alpha}^0 + \tilde{\beta}^0$$

By partitioning the multinomial Dirichlet into binomial Beta distributed estimates, the easy interpretability of the binary representation is maintained. By interactively choosing from a given set of relations and categorical partitions of the Dirichlet in the Histogram, the flexibility of the histogram representation is maintained. For more complex evaluations and combinations of the resulting different binomial *CertainTrust* opinions, its belief logic extension, *CertainLogic* [175], is also available.

3.2.9 *Section Summary*

This section has introduced an extension of the *CertainTrust* model, extending it from the binomial to the multinomial. As in Section 3.1, the main focus was on a sound statistical footing, particularly with regard to certainty estimation. To this end, both the trust estimator and the certainty estimators were formally derived from the multinomial distribution. The certainty estimators introduced for the binomial case serve as a basis to also provide accurate certainty estimates in the multinomial case. Applying statistical theory, multinomial certainty estimators are constructed from simultaneous credibility and confidence intervals, resulting in the Simultaneous Credibility Interval-based Certainty Estimator for Multinomial Proportions and the Simultaneous Confidence Interval-based Certainty Estimator for

Multinomial Proportions. As for the binomial certainty estimators in Section 3.1, the adherence of the multinomial certainty estimators to the properties postulated by Wang & Singh [196] is shown.

Furthermore, the parameters for initial trust values are related to Bayesian priors and the choice of non-informative priors for the multinomial case is briefly discussed. Finally, an extension of the HTI to the multinomial case is suggested.

3.3 CHAPTER SUMMARY

In this chapter, the statistics behind the *CertainTrust* model were motivated, revised and extended. In particular, the certainty estimation was given a new interpretation, formally derived from binomial and categorical distributions underlying the binomial and multinomial case, respectively. This interpretation considers certainty an estimate of the dispersion of the trust score computed in *CertainTrust*. Fundamentally, this also advances the state-of-the-art presented in works by Wang & Singh [196] and Teacy et al. [189], leverages proven statistical methods (see, e.g., [27]), and provides a flexible extension from the binomial into the multinomial case of trust assessment.

In the second part of this Chapter, the binomial *CertainTrust* model was extended to the *Multinomial CertainTrust* model, providing simultaneous confidence interval-based certainty estimators and graphical representations of the results via an integration of histogram-style bar charts and the HTI.

Specific contributions in this chapter include:

- For the *binomial case* of trustworthiness assessment:
  - *Credibility Interval-based Certainty Estimator*, Definition 10, p. 59: A dispersion-based certainty estimator, derived from the Bayesian Jeffreys credibility interval for binomial proportions.
  - *Confidence Interval-based Certainty Estimator*, Definition 12, p. 64: A dispersion-based certainty estimator, derived from the frequentist Wilson confidence interval for binomial proportions; providing a closed-form alternative to the open-form Credibility Interval-based Certainty Estimator at comparable performance levels.
  - An adjusted computation of the *CertainTrust* expectation value, in order to incorporate the novel certainty estimators into the *CertainTrust* model (Section 3.1.7).
  - An augmented HTI capable of displaying the potential dispersion of a trust estimate.
- For the *multinomial case* of trustworthiness assessment:

- *Multinomial CertainTrust*: A complete extension of the predictive model behind *CertainTrust* to handle multinomial opinions
- *Simultaneous Credibility Interval-based Certainty Estimator for Multinomial Proportions*, Definition 19, p. 83: A version of the Credibility Interval-based Certainty Estimator that corrects for the multiple testing inherent in multinomial proportions.
- *Simultaneous Confidence Interval-based Certainty Estimator for Multinomial Proportions*, Definition 20, p. 85: A closed-form alternative to the Simultaneous Credibility Interval-based Certainty Estimator, using Goodman's correction of the Wilson confidence interval.
- A mapping of multinomial priors to *Multinomial CertainTrust* initial trust parameters and corresponding *MultinomialCertainTrust* expectation value computation (Sections 3.2.6 and 3.2.7).

The contributions of this chapter have increased the statistical soundness of the binomial estimation model underlying *CertainTrust*. Especially the certainty estimation has been improved by using credibility/confidence intervals as a basis for computing a certainty score. By doing so, the certainty estimate uses the available information on the variance of the estimated parameter to give a more exact estimate of the exactness of the trustworthiness estimate. Furthermore, the certainty estimate can be scaled by the user by varying the credibility or confidence parameter z of the interval. This value is a standard statistical term and controls the remaining uncertainty that a user is willing to accept. Additionally, *CertainTrust* has been considerably extended into a multinomial model, capable of handling fine granular feedback, without sacrificing statistical soundness. In combination, these advances permit a more exact estimation of trust and uncertainty, for instance, in industrial applications.

Overall, this chapter has provided a complete prediction model for binomial *and* multinomial trustworthiness assessment, representing the core of a more comprehensive trust model. Increasing the comprehensiveness of the core model, in the following Chapter 4, methods for trust information processing are considered. This includes mechanisms for trust propagation and determining recommender trustworthiness, as well as detecting and dealing with changes in trustee behaviour.

4

TRUSTWORTHINESS INFORMATION PROCESSING

In order to provide a comprehensive computational model of trust, *CertainTrust* [173] provides operations for combining and aggregating the direct observations made by the truster with recommendations from third parties. The operators in *CertainTrust* extend the fundamental concepts presented in the Beta Reputation System [108], thereby enabling *trust propagtion* in a conceptual trust overlay network. In this Chapter, a number of contributions with regard to the processing of opinions for trust propagation are presented. This includes the extension of mechanisms for combining opinions (discounting, consensus), averaging opinions (fusion), and determining recommender trustworthiness, as well as the introduction of change point detection methods for replacing and augmenting current approaches used in the ageing of opinions.

The *consensus* and *discounting* operations are extended to the *Multinomial CertainTrust* model; additionally, a variation and extension of the *fusion* operation for averaging is presented, which was originally introduced in and is further adapted from [77].

A central theme in this chapter is the application of statistical hypothesis tests; these tests can be used for determining recommender trustworthiness, for extending the conflict-aware fusion operation and within the scope of change point detection mechanisms. The frequentist[1] *Fisher's Exact Test* (FET) [56] and its multinomial extension, the *Fisher-Freeman-Halton Test* [58], are leveraged in order to provide probability estimates on the independence of opinions. These probability estimates form the basis for determining the degree of similarity when computing recommender trustworthiness, the degree of conflict in the conflict-aware fusion operation, and the presence of a change in trustee behaviour.

Section 4.1 briefly introduces the concept of trust propagation through recommendations. Section 4.2 introduces various methods for computing the trustworthiness of recommenders and extends the state of the art in two directions: first, it introduces a non-sequential test-based method that accounts for the dependence of subsequent observations when receiving recommendations from one particular recommender; second, it provides the capabilities for dealing with multinomial recommendations. Section 4.3 presents extensions to the consensus, discounting and fusion operations required for trust propagation and shows how they can be used to determine recommender trust-

1 For the presented application, the frequentist and Bayesian approaches coincide. In general, an adaptation of Bayes factors [116] for use in computational trust assessment presents itself as a feasible direction for future work.

worthiness as a discounting factor. Section 4.4 discusses the use of ageing and change point detection for dealing with non-stationarity of trustee behaviour[2].

The differentiation between binomial and multinomial estimation is not as stringent in this chapter as it was in Chapter 3. Except for the discounting, consensus and fusion operations, which are explicitly extended to the multinomial case, the other methods presented in this chapter leverage the FET. All of these methods are presented for the binomial case. However, by substituting the FET with the *Fisher-Freeman-Halton Test* when computing recommender trustworthiness and computing the degree-of-conflict in conflict-aware fusion, multinomial data can be processed in these two cases. Similarly, the change point detection method by [178] that is applied to trustworthiness estimation in Section 4.4 is illustrated in the binomial case but can also be applied to multinomial data.

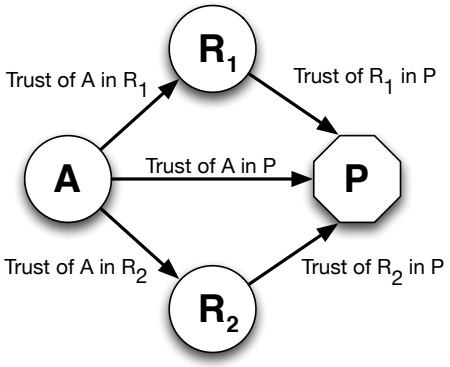


Figure 5: Trust network (see also [173])

.

4.1 TRUST PROPAGATION, ROLES AND CONTEXT

Trust propagation is the process of sharing trust-related information over a *trust network* in the form of *recommendations*. Figure 5 depicts a very simple conceptual trust network, consisting of four abstract *entities* that act in three distinct and different *roles*.

In this example, entity A fills the role of *truster*, evaluating a potential partner in an interaction, P, which therefore takes on the (passive) role of *trustee*. A evaluates the trustworthiness of P, based on its own observations of the past behaviour of entity P, *as well as* recommendations A receives from entities R_1 and R_2, acting as *recommenders*.

Following common convention (see, for example, [173]), interactions and recommendations are recorded within a specific *context*, C. Such a context may be the provisioning of a particular virtual (for

2 In the related work on computational trust, this is sometimes referred to as *dynamism*.

instance, a cloud storage service) or real-world service (for instance, a car repair service). Within each context $\mathcal{C}$, two particular types of sub-context can be distinguished: the sub-context describing the performance of the interaction partners within an interaction, $\mathcal{C}(I)$, and the sub-context of recommender performance within the given context, $\mathcal{C}(R)$.

Within context $\mathcal{C}$, entity P is evaluated with regards to an expected quality of service. If selected by entity A, when P interacts with A, P gains or loses trust in the eyes of A within sub-context $\mathcal{C}(I)$. In a manner of speaking, entities R_1 and R_2 also provide a service in context $\mathcal{C}$ to entity A by providing recommendations. They gain trust with A based on the accuracy of their recommendations. While the evaluation criteria of service provisioning between truster and trustee are generally beyond the purview, the recommendation mechanism forms an integral part of the trust model. Therefore, the evaluation criteria of *recommendation* provisioning are have to be considered. This is of particular interest, as the trustworthiness of the recommenders in context $\mathcal{C}(R)$ is used to weight the recommendation. As already briefly introduced in Chapter 2.1.2, one should keep in mind that trustworthiness estimation is dependent on the context in which an interaction takes place. Thus, an interactor might prove trustworthy in one particular context (such as providing musical entertainment) and not in another (such as making good business decisions). Similarly, a recommender may be competent in giving recommendations in a particular context (such as recommending a particular recording artist) and not in another (such as recommending a particular financial investment advisor). However, as the estimation and processing *mechanisms*, such as trust and certainty estimators, do not vary across contexts, the explicit declaration of a specific context is generally omitted in this thesis. It suffices to state that trustworthiness estimation occurs under an arbitrary but fixed context, i.e., $\mathcal{C}$. The following Sections will be concerned with the processing mechanisms enabling trust propagation.

Specifically, the focus of Section 4.2 will be on determining the trustworthiness of recommenders in trust propagation. In Section 4.3, three different operations will be introduced and extended from their binomial *CertainTrust* [173] original forms. Where deemed necessary, further statistical methods will be applied. These three operations are:

- *Discounting*: This operation weights observations provided by a recommender according to the trustworthiness of said recommender.
- *Consensus*: The consensus operation combines observations *made independently of each other* by the truster and its recommenders. This means that the individual observations have been made for distinct and independent events, thereby satisfying the assumption of statistical independence of the resulting, combined sam-

ple. Since the combined sample includes the past observations of A *and* the recommenders, the amount of evidence available for making an estimate increases.

- *Fusion*: The fusion operation aggregates observations by averaging opinions for which the independence assumption *cannot* be guaranteed to hold. This operation is not part of the original *CertainTrust* model, but is derived from *Subjective Logic* [103]. It has been applied to and made compatible with *CertainTrust* opinions in [175]. The weighted and conflict-aware variation used (and extended) in the following has been presented for binomial opinions in [77], to which the author has contributed to the mathematical formulation of the weighting and conflict-awareness computations.

It should be noted, that the mechanisms behind trust propagation presented in *CertainTrust* [173] and *Subjective Logic* [104] result in opinions that are technically not Beta distributed anymore, although they are treated as such [153]. Muller and Schweitzer [153] provide a formal treatment on this and trust chains for propagation in Beta models. However, the representation chosen for *CertainTrust* [173] and *Subjective Logic* [104] is a reasonable and practicable approximation.

4.2 RECOMMENDER TRUSTWORTHINESS

A key element of trust propagation is the estimation of recommender trustworthiness for use in the discounting and consensus operations when compounding opinions from direct observations and recommendations. The estimation of recommender trustworthiness bears a strong resemblance to the general trustworthiness estimation presented before – with a number of marked differences. The estimation procedures presented in Chapter 3 rely on a set of assumptions, among them:

- independent, identically distributed (*iid*) random variables,
- discrete n-ary, exhaustive and mutually exclusive observations, and
- conjugate Dirichlet priors (with Beta priors being a special case of the Dirichlet for binary observations)

Considering the way that a recommendation service is provided, a number of caveats arise when determining the trustworthiness of the recommender. Consider an entity, A, evaluating the trustworthiness of a population of potential interaction partners, $P_1, P_2, \ldots, P_m$ within the same, fixed context. Aside from its own observation, entity A can additionally rely on a population of recommenders, $R_1, R_2, \ldots, R_k$. Thus, at any point in time a recommender, R_i, can give at most as

many recommendations as there are potential interaction partners, that is, $m \in \mathbb{N}$. Of these recommendations only those regarding a (potential) interaction partner P_j that at some point have actually led to an interaction between A and P_j can be evaluated with regards to their accuracy – in terms of the deviation of what the R_i recommended and what A itself has experienced. That is, not every recommendation leads to an observation that can confirm or refute the recommendation's accuracy. And for those that lead to an observation, a measure of recommendation accuracy has to be selected. This can be a discrete classification of the accuracy of the recommendation or a continuous residual value.

Furthermore, assume that over the course of time recommender R_i is a asked by A to provide a recommendation on the *same* interaction partner P_j at different points in time. The recommender, R_i would report its own experience with P_j. R_i reports its opinion on P_j based on its own past observations, either as a *CertainTrust* opinion (Section 3.2.1, p. 77), or the concordant sufficient statistics (for instance, the sums of successes and failures in a binomial sample). Then, the two recommendations of R_i on P_j are generally not independent of each other. For example, consider a recommendation by R_i on P_j at time z consisting of the sum of success, r, and the sum of failures, s, and let it be $(r = 16, s = 2)_z$. Suppose that at a later time, $z+$, R_i has had six additional interactions with P_j, four success and two failures. Its second recommendation to A on P_j at time $z+$ would then be $(r = 20, s = 4)_{z+}$.

Obviously, the second recommendation contains all 18 observations of the first recommendation and does not contain new information *exclusively*. The first opinion informs the second one to a considerable degree, so that the value of the second *depends* on the first. Therefore, the new recommendation does not form the basis for a new, *additional* observation on the trustworthiness of recommendation by recommender R_i, but rather *supersedes* older information. Therefore, while the ability of A to assess the accuracy of R_i's recommendation on P_j improves over time, the additional information gained to assess the trustworthiness of R_i's general ability as a recommender increases only incrementally. Therefore, an individual estimate of the reliability of recommendations of a specific recommender R_i and a specific trustee P_j has to be maintained. Over all these estimates for a particular recommender, R_i, the overall trustworthiness of that recommender can then be determined, in a *second* step.

Without even considering the specific measure of accuracy used by A to determine a recommender's trustworthiness, it can already be seen that the task of establishing the trustworthiness of recommenders is slightly different from the general trustworthiness assessment discussed heretofore. This is further compounded by the way

the accuracy of a recommendation by R_i is determined by truster A after an interaction with trustee P_j.

In the original *CertainTrust* model [173], two ways of updating trust information on recommenders are introduced. One is based on a classification scheme considering only the last interaction, the other on computing residuals between the recommendations estimate and the estimate based on direct observations by A. They will be briefly introduced and their respective shortcomings will be outlined.

For this, we will use the nomenclature for opinions used in the corresponding chapters in [173]. As the original *CertainTrust* model only provides trust assessment in the binomial case, we will for now consider binomial opinions exclusively. Let $o^A_{R_i} = (r^A_{R_i}, s^A_{R_i})^{rs}$ be an opinion of entity A on recommender R_i reporting the sufficient statistics *sum of successes*, $r^A_{R_i}$, and *sum of failures*, $s^A_{R_i}$. The subject of such an opinion is, for the reasons outlined above, not the overall accuracy of the trustworthiness estimate of recommender R_i in the eyes of A, but only the the accuracy with respect to a specific trustee P_j. Therefore, this information will be amended to the – admittedly already cluttered – declaration, so that: $o^A_{(R_i,P_j)} = (r^A_{(R_i,P_j)}, s^A_{(R_i,P_j)})^{rs}$.

4.2.1 *Tendency Classification Update Considering only the Last Interaction*

Suppose recommender R_i has given a recommendation on P_j to trustee A. A has subsequently selected trustee P_j for an interaction, has interacted with the trustee and graded the interaction as either positive or negative. Let $\omega^{R_i}_{P_j} = (t^{R_i}_{P_j}, c^{R_i}_{P_j})$ be the binomial *CertainTrust* opinion that R_i has supplied to A as a recommendation. Furthermore, let the variable fb be the feedback grading the interaction between A and P_j, so that $fb = 1$ if the interaction was successful, and $fb = -1$ if the interaction was unsuccessful. When considering only the last interaction, the update of A's opinion on R_i, according to [173], is as follows:

- if $(2 \cdot t^{R_i}_{P_j} - 1) \cdot fb > 0$, then the updated opinion of A on R_i is $o^A_{(R_i,P_j)} = (r^A_{(R_i,P_j)} + 1, s^A_{(R_i,P_j)})^{rs}$,
- if $(2 \cdot t^{R_i}_{P_j} - 1) \cdot fb < 0$, then the updated opinion of A on R_i is $o^A_{(R_i,P_j)} = (r^A_{(R_i,P_j)}, s^A_{(R_i,P_j)} + 1)^{rs}$
- if $(2 \cdot t^{R_i}_{P_j} - 1) \cdot fb = 0$, then the updated opinion of A on R_i is $o^A_{(R_i,P_j)} = (r^A_{(R_i,P_j)}, s^A_{(R_i,P_j)})^{rs}$

In less formal terms, if the recommendation $\omega^{R_i}_{P_j}$ indicates that the next interaction between A and P_j is going to be a positive on, i.e., $t^{R_i}_{P_j} > 0.5$, this classification into the class '*positive*' is taken as the sole information contained in the recommendation. Should the feedback

grading the interaction also be positive, the classification by R_i is considered accurate and the sufficient statistic *sum of successes*, $r^A_{(R_i,P_j)}$ is incremented accordingly. If the classification was '*negative*' and the feedback grading confirmed the classification by also being negative, the mechanism behaves likewise. If, however, the feedback grading and the classification are not identical, meaning that R_i has supposedly misclassified the interaction, the sufficient statistic *sum of failures*, $s^A_{(R_i,P_j)}$ is incremented instead. In case the recommendation was inconclusive, i.e., $t^{R_i}_{P_j} - 1) = 0.5$, none of sufficient statistics is incremented.

The classification approach for updating the opinion on a recommender sacrifices information by neglecting both the actual probability estimate contained in the opinion, $t = \hat{p}$, as well as the certainty estimate c that is reported by the recommender. Additionally, it is not compared to the complete history of observations that truster A has compiled on P_j and does not take into account that repeat recommendations by R_i on P_j are not generally statistically independent. It does however maintain the discreteness assumptions underlying the application of a Beta distribution used in the estimation.

4.2.2 *Linear Update Estimation Considering the Interaction History*

In order to leverage the information available for the estimation more efficiently, [173] proposes a second update mechanism for recommender trustworthiness[3]. For a recommendation by R_i on P_j, this approach takes into account the observations that A has made with regards to P_j, the certainty, $c^{R_i}_{P_j}$, that R_i reports in its recommendation, as well as the value of $t^{R_i}_{P_j}$. Let $t^A_{P_j}$ be the trust estimate that truster A can establish on P_j after the current interaction, relying solely on A's own observations resulting from direct interactions between itself and P_j. Then, the absolute residual between R_i's recommendation and A's observations can easily be computed as $|t^A_{P_j} - t^{R_i}_{P_j}|$ and used in the update. The update procedure proposed in [173] is as follows:

$$r_{new} = c^{R_i}_{P_j} \cdot (1 - |t^A_{P_j} - t^{R_i}_{P_j}|)$$
$$s_{new} = c^{R_i}_{P_j} \cdot (|t^A_{P_j} - t^{R_i}_{P_j}|)$$

so that the updated opinion of A on R_i (with regard to recommendations on P_j) is

$$o^A_{(R_i,P_j)} = (r^A_{(R_i,P_j)} + r_{new}, s^A_{(R_i,P_j)} + s_{new})^{rs}$$

Notice how the updated opinion is put into the framework of a binary proportion estimation task, as though it were part of a binomial sample. This would *imply iid*, discrete observations in two

3 Their approach is largely identical to the one proposed in [108].

mutually exclusive categories. However, the residuals and hence the observations in case of repeat recommendations from R_i on P_j are not *iid*. Additionally, the residuals are continuous variables in $[0;1]$ and the update mechanism violates the condition of mutual exclusivity. Nonetheless, certainty estimates are supposed to assume Beta distributed posteriors.

The goal of the estimation of recommender trustworthiness thus becomes determining the average absolute residual of generally not independent, continuous observations. Clearly, the distributional assumptions permitting the use of a Beta prior do not hold any longer, and another prior distribution should be chosen. This, however, requires a new assumption on the distribution of the residuals. Possibly, Gaussianity could be assumed for a reasonably large number of observations and a truncated Normal distribution [101] can be applied. The truncation would have to take place to account for the fixed-length carrier $[0;1]$, while the use of a Normal distribution might be warranted by the application of the central limit theorem.

Even so, the fact that the observations may not be *iid* theoretically requires a more complex estimator that takes dependence into account. Such estimators, in the framework of maximum likelihood estimation, have been introduced, for instance, in [38, 88].

Due to the simple nature of the dependence between the individual recommendations and their consistency as an estimate, the resulting estimate for the absolute residual remains consistent as well. Thus, purely as an estimator for the trustworthiness of recommender $R_{i,j}$, based on maximum likelihood estimation, the approach proposed by Ries in [173] is a reasonable heuristic. The caveat concerns the assumption of independence and the implication of a Beta distribution for making further inferences, such as statistically meaningful certainty estimates.

4.2.3 *Further Sequential Update Rules for Recommender Trustworthiness*

Variations of the sequential recommender trustworthiness update process are in widespread use in the related work. Wang et al. [198] provide an overview that will be briefly summarised here.

- *Jøsang/Linear-WS*: The recommender trust update mechanism in [108] is identical to the one presented above in Section 4.2.2.
- *Max-Certainty*: The max-certainty approach [198] has the following update rule

$$r_{new} = c_{P_j}^{R_i} \cdot \frac{(t_{P_j}^{R_i})^{r_{P_j}^{A}} \cdot (1 - t_{P_j}^{R_i})^{s_{P_j}^{A}}}{(t_{P_j}^{A})^{r_{P_j}^{A}} \cdot (1 - t_{P_j}^{A})^{s_{P_j}^{A}})}$$

$$s_{new} = 1 - r_{new}$$

- *Sensitivity*: The recommender trust update mechanism introduced in [189], called *sensitivity* by [198]

$$r_{new} = c_{P_j}^{R_i} \cdot \frac{(t_{P_j}^{A})^{r_{P_j}^{R_i}} \cdot (1 - t_{P_j}^{A})^{s_{P_j}^{R_i}}}{(t_{P_j}^{R_i})^{r_{P_j}^{R_i}} \cdot (1 - t_{P_j}^{R_i})^{s_{P_j}^{R_i}})}$$

$$s_{new} = 1 - r_{new}$$

- *Average*-β: Here, an average *accuracy* argument is used to update recommender trust, as introduced in [198]

$$q = \sqrt{\left(t_{P_j}^{A} - \frac{r_{P_j}^{R_i} + 1}{r_{P_j}^{R_i} + s_{P_j}^{R_i} + 2}\right)^2 + \frac{(r_{P_j}^{R_i} + 1) \cdot (s_{P_j}^{R_i} + 1)}{(r_{P_j}^{R_i} + s_{P_j}^{R_i} + 2)^2 \cdot (r_{P_j}^{R_i} + s_{P_j}^{R_i} + 3)}}$$

$$r_{new} = c_{P_j}^{R_i} \cdot c_{P_j}^{A} \cdot (1 - q)$$

$$s_{new} = c_{P_j}^{R_i} \cdot c_{P_j}^{A} \cdot q$$

The update is incremental for all of the update approaches listed, so that the updated opinion of A on R_i (with regard to recommendations on P_j) is

$$o_{(R_i,P_j)}^{A} = (r_{(R_i,P_j)}^{A} + r_{new}, s_{(R_i,P_j)}^{A} + s_{new})^{rs}$$

All of these sequential update mechanisms are *heuristics* for determining recommender trustworthiness. While some offer very good predictive performance (see Section 4.3.6), in particular the Linear and Average-β approaches, they still suffer from a number of shortcomings from a theoretical point of view, such as: An abuse of the Beta distribution as a conjugate prior[4], and a difficult extension of the measures to the multinomial case of trustworthiness estimation. Additionally, considerations of statistical dependence of the observations are not specifically taken into account.

However, if repeat sampling and sequential update can be avoided altogether by reformulating the problem, complicating the estimation by assuming dependence can be foregone and an easy extensibility to multinomial recommendations exists. A novel method for determining recommender trustworthiness is therefore introduced in the following. This method utilises a well-known *exact statistical test* and does not rely on sequential update, but on hypothesis testing at regular intervals.

4 From a practical point of view, this does not impact the probability estimation, as the MLE for the first moment is the same for Beta and Gaussian distributions.

4.2.4 *Exact Test-based Recommender Trustworthiness*

As we have seen from the previous sections, estimating the trustworthiness of a recommender means determining the similarity between the recommendations reported by that recommender and the direct observations made by the truster. From a modelling perspective, both the recommender and the truster are observing Bernoulli or Categorial random processes. Therefore, the estimation task can be reduced to estimating the probability that both the recommendation and the observations from direct interactions made by the truster have been generated by the *same*[5] random process. Thus, for Bernoulli random processes, estimating recommender trustworthiness means estimating the probability that $\mathrm{Bin}(n, p_A) = \mathrm{Bin}(m, p_{R_i})$, where p_A is the parameterisation of the probability of success of the process observed by truster A, and p_{R_i} the corresponding parameterisation of the process observed by recommender R_i.

The recommendation of recommender R_i on P_j is reported to A as an opinion $o_{P_j}^{R_i} = (r_{P_j}^{R_i}, s_{P_j}^{R_i})^{rs}$, where $r_{P_j}^{R_i}$ and $s_{P_j}^{R_i}$ are the sufficient statistics *sum of successes* and *sum of failures* in previous interactions between R_i and P_j. From its past observations of interactions with P_j, the truster A holds its own, direct opinion of P_j, $o_{P_j}^{A} = (r_{P_j}^{A}, s_{P_j}^{A})^{rs}$. Estimating the probability that the opinions $o_{P_j}^{R_i}$ and $o_{P_j}^{A}$ originate from observing the same random process is equivalent to estimating the probability that recommender R_i reported correctly on the trustworthiness of P_j, given what A knows itself of the trustworthiness of P_j. Thus the goal of the estimation is to determine the probability of independence of $o_{P_j}^{R_i}$ and $o_{P_j}^{A}$, which will be construed as:

$$p(p_A = p_{R_i}; r_{P_j}^{A}, s_{P_j}^{A}, r_{P_j}^{R_i}, s_{P_j}^{R_i})$$

In order to clarify the notion of *independence* of $o_{P_j}^{R_i}$ and $o_{P_j}^{A}$ under the given circumstances, assume that $o_{P_j}^{R_i}$ and $o_{P_j}^{A}$ were generated by two different random processes with different probabilities of success:

- one observed by the recommender, R_i, with probability of success p_{R_i}, and reported in $o_{P_j}^{R_i}$, and
- another observed by the truster, A, with probability of success p_A, and reported in $o_{P_j}^{A}$.

Now, the opinions are *dependent* on the entity reporting it. If they were independent of the entity reporting them, the two processes would likely have generated two indistinguishable opinions, which, for a sufficiently large number of observations, would indicate that $p_A \approx P_{R_i}$. Thus, in the latter case of independence, it does not matter

5 Here, *'same'* refers to the parameterisation of the random process.

whether the opinions were reported by A or R_i, as the random process generating them can be assumed to be identical.

In the classic statistics literature, a number of statistical *tests for independence* exist[6], such as the well known χ^2-*Test* [55] or the *Wilcoxon/Mann-Whitney Test* [203]. For the test for independence of two Binomial samples, expressed in $o^{A}_{P_j}$ and $o^{R_i}_{P_j}$ as the sufficient statistics $r^{P_j}_{A}$, $s^{P_j}_{A}$, $r^{P_j}_{R_i}$ and $s^{P_j}_{R_i}$, *Fisher's Exact Test* (FET) exists. This test is exact, as it does not rely on Gaussian approximations, and is therefore applicable to small sample sizes, unlike the χ^2-*Test*. The FET relies on the *sample odds ratio* (Definition 25) to compute the probability that the two samples were generated from the same Bernoulli process.

In addition, an extension of the FET, the *Fisher-Freeman-Halton Test* [58][7], generalises the test so that it is applicable to multinomial opinions. Thus, the method for determining recommender trustworthiness can also be used in the multinomial case of trustworthiness assessment.

The following three definitions, Definition 24, p. 107 to Definition 26, p. 108, introduce the odds ratio and *Fisher's Exact Test*. For the FET, the null-hypothesis, h_0: *the two samples are independent*, translates directly to h_0: *the sample odds ratio of the two samples equals 1*. The resulting p-value reported by the FET represents the probability that h_0 cannot be rejected, and thus that the samples are not dependent.

Definition 24 (Odds Ratio [3])**.** Let p be the *probability of success* in a binomial sample, then the *odds of success* are defined to be

$$\text{odds} = \frac{p}{1-p}$$

Let p_1 and p_2 be the probability of success of two binomial samples. Then the *odds ratio* between the two samples is defined as

$$\text{OR} = \frac{\text{odds}_1}{\text{odds}_2} = \frac{\frac{p_1}{1-p_1}}{\frac{p_2}{1-p_2}} = \frac{p_1 \cdot (1-p_2)}{p_2 \cdot (1-p_1)}$$

Definition 25 (Sample Odds Ratio [3])**.** The *sample odds ratio* is an estimator of the odds ratio based on observed frequencies. For two samples, it equals the ratio of the sample odds in the two samples. Let r_1 be the *sum of successes* and s_1 be the *sum of failures* in the first sample, r_2 and s_2 the corresponding sums for the second sample. Then the sample odds ratio is defined as

$$\widehat{\text{OR}} = \frac{\frac{r_1}{s_1}}{\frac{r_2}{s_2}} = \frac{r_1 \cdot s_2}{r_2 \cdot s_1}$$

6 see, for instance, [3]

7 The extended version of Fisher's Exact Test by Freeman and Halton [58] is only one (early) example, various other tests, based on different statistics, exist, see [2].

A sample odds ratio of $\widehat{OR} = 1$ *indicates* that two samples are in fact generated by the same process; this is leveraged in the FET, as defined in the following:

Definition 26 (Fisher's Exact Test (FET) [56])**.** Let $X = (x_1, x_2, \ldots, x_n)$ and $Y = (y_1, y_2, \ldots, y_m)$ be two sequences of realisations of Bernoulli distributed random variables X_i and Y_j. Let r_X, r_Y be realisations of the sufficient statistics R_X and R_Y, the *sum of successes* in X and Y respectively. Let s_X and s_Y the corresponding realisations of the *sum of failures*, S_X and S_Y. Furthermore, let the null-hypothesis, h_0, be that X and Y are conditionally independent[8], that is, they were generated from the *same* Bernoulli process. The alternative hypothesis, h_a would then be that X and Y are *not* conditionally independent. This relates to the sample odds ratio $\widehat{OR}$, as h_0 can formulated as

$$h_0 : \widehat{OR} = 1$$

Leveraging an algebraic argument, the probability of observing a particular realisation of r_X, r_Y, s_X and s_Y is given by a hypergeometric distribution:

$$p(r_X, r_Y, s_X, s_Y) = \frac{\binom{r_X+s_X}{r_X}\binom{r_Y+s_Y}{r_Y}}{\binom{r_X+r_Y+s_X+s_Y}{r_X+r_Y}}$$

To test h_0: *independence*, the p-*value* of Fisher's Exact Test (FET) is the sum of hypergeometric probabilities for realisations of R_X, R_Y, S_X, S_Y that are at least as favourable to the alternative hypothesis h_a as the observed realisation, under fixed marginal sums $R_X + S_X = r_X + s_X$, $R_X + R_Y = r_X + r_Y$, $S_X + S_Y = s_X + s_Y$ and $R_Y + S_Y = r_Y + s_Y$.

Based on the FET, a novel method for estimating recommender trustworthiness can now be defined by combining the p-value of the FET with a certainty estimate and computing the *Certain Trust* expectation value ([175] and Section 3.1.7, p. 69) from the p-value and certainty estimate.

Definition 27 (Fisher's Exact Test/FET-based Recommender Trustworthiness Estimator)**.** Let $o_{P_j}^{R_i} = (r_{P_j}^{R_i}, s_{P_j}^{R_i})^{rs}$ be a recommendation from recommender R_i on trustee P_j to truster A, and $o_{P_j}^{A} = (r_{P_j}^{A}, s_{P_j}^{A})^{rs}$ be the corresponding opinion of A on R_i from past direct interactions between A and R_i. Let $\hat{p} = t_{R_i}^{A}$ be the p-value returned by Fisher's Exact Test when testing for independence of $o_{P_j}^{A}$ and $o_{P_j}^{R_i}$. Furthermore, let $c_{P_j}^{A}$ and $c_{P_j}^{R_i}$ be the certainty values of opinions $o_{P_j}^{A}$ and $o_{P_j}^{R_i}$.

Then the *Fisher's Exact Test/FET-based Recommender Trustworthiness Estimate* is defined as the *CertainTrust* opinion $\omega_{R_i}^{A} = (t_{R_i}^{A}, \min(c_{P_j}^{A}, c_{P_j}^{R_i}), f)$ with expectation value

$$E(\omega_{R_i}^{A}) = \min(c_{P_j}^{A}, c_{P_j}^{R_i}) \cdot \hat{p} + (1 - \min(c_{P_j}^{A}, c_{P_j}^{R_i})) \cdot f$$

8 Here, dependence and independence refer to the entities (truster or recommender) making the observations; if the probabilities of success of the two Bernoulli processes are identical, the sequences do not depend on the entity making the observation.

Computing the *Fisher's Exact Test/FET-based Recommender Trustworthiness Estimate* for a particular recommender does not require a sequential update. Rather, old estimates are superseded or replaced by the new estimate. It thus avoids any theoretical concerns of updating with highly dependent observations.

The certainty estimates $c^{A}_{P_j}$ and $c^{R_i}_{P_j}$ are computed separately for the two opinions $o^{A}_{P_j}$ and $o^{R_i}_{P_j}$, ideally using a statistically derived certainty estimator, such as the *Wilson Interval Certainty Estimator* presented in Definition 12, p. 64. The statistical power of the FET is dependent on the sizes of the two samples it is given as input. In particular, the smaller of the two samples determines how reliable the FET is. Therefore, the smaller of the two certainty estimates $c^{A}_{P_j}$ and $c^{R_i}_{P_j}$ is used to estimate the certainty of the trustworthiness score computed by the FET-based Recommender Estimator. For the initial trust value f, a non-informative prior will be assumed in the following, leading to $f = 0.5$.

As a further consideration, it should be noted that the exact test statistic used in the FET can become prohibitively expensive to compute for large values of the *sums of success*, $r^{A}_{P_i}$ and $r^{R_i}_{P_i}$, and the corresponding *sums of failures*, $s^{A}_{P_i}$ and $s^{R_i}_{P_i}$. However, in this case it is reasonable to substitute the FET with an approximate statistic, such as the χ^2, as this test's Gaussianity assumptions typically hold under large sample sizes, thereby maintaining the feasibility of the general approach.

4.2.4.1 *Recommender Trustworthiness from Multinomial Recommendations*

One major advantage of using a statistics hypothesis test for computing recommender trustworthiness is its easy application to the multinomial case. Given two *multinomial opinions* of the same length, instead of the binomial opinions discussed before, the extended Fisher's Exact Test [58] (or *Fisher-Freeman-Halton Test*) will return as its p-value a probability estimate of the independence of the two opinions, just as in the binomial case.

In the multinomial case, the corresponding multinomial opinions may provide a multinomial certainty estimate, such as the one given by the *Goodman Interval Certainty Estimator for Multinomial Proportions* (Definition 20, p. 85). This can be made compatible with the definition of the FET-based Recommender Trustworthiness Estimator (Definition 27), by first selecting the minimum element of each multinomial certainty vector. Thus, if $\vec{c}^{A}_{P_j}$ is the vector of certainties reported for each of the multinomial proportions in a multinomial opinion $o^{A}_{P_j} = (\alpha_1, \alpha_2, \ldots, \alpha_m)^{\alpha}$, and $\vec{c}^{R_i}_{P_j}$ the corresponding vector for a second opinion $o^{R_i}_{P_j} = (\alpha'_1, \alpha'_2, \ldots, \alpha'_m)^{\alpha}$, this yields

$$\omega^{A}_{R_i} = (t^{A}_{R_i}, \min(\min(\vec{c}^{A}_{P_j}), \min(\vec{c}^{R_i}_{P_j})), f)$$

The hypothesis test used in the multinomial case is fundamentally identical to the one presented above for the binomial case, except for its application to m-dimensional samples. However, other tests may be substituted in order to account for the nature of the categories in the multinomial sample. In case of ordered categories – that is, ordinal data in the sample – the efficient score test by Agresti, Mehta and Patel [5] may provide better performance for multinomial samples [2]. However, the principle of applying the *Test-based Recomender Trustworthiness Estimator* remains unchanged.

It should be noted that the question of whether a recommender provides trustworthy recommendations within context $\mathcal{C}$ is considered as a binomial problem. Thus, the opinion of entity A on recommender R_i will always be a binomial one. This is irrespective of the multinomial nature of the recommendations and opinions held on the potential trustee. Thus, if $o_{P_j}^{A} = ((\alpha_1, \alpha_2, \ldots, \alpha_m)^{\alpha}$ is the opinion of A on trustee P_j, and $o_{P_j}^{R_i} = (\alpha'_1, \alpha'_2, \ldots, \alpha'_m)^{\alpha}$ is the recommendation from R_i on P_j, both multinomial opinions of dimension $m \in \mathbb{N}, m > 2$, the resulting opinion of A on the trustworthiness of R_i in recommending P_j is still a binomial opinion. This is motivated by the desired use of the recommender trustworthiness estimate as the basis for a *discounting* factor in trust propagation.

4.2.5 *Section Summary*

In this section, methods for estimating the trustworthiness of recommenders in trust propagation have been presented. After reiterating sequential update methods from the related work, a novel method based on *Fisher's Exact Test* [56] has been introduced. The *FET*-based recommender trustworthiness uses a statistical argument to compute the probability that two samples, that is, the recommendations given by a recommender and the direct experience by the trustee itself, are generated from the same statistical process. By avoiding sequential update mechanisms, the novel *FET*-based method does not have to compensate for the statistical dependence of consecutive recommendations by the same recommender; additionally, it can be computed at any time from the current history of direct experiences held by the trustee and a current recommendation by a recommender, thereby eliminating the need to constantly monitor recommender performance at every time step. Section 4.3.6, p. 126, shows performance comparisons of the different recommender trustworthiness estimation methods and evaluates their efficacy.

The p-value reported by *Fisher's Exact Test* represents an actual, exact probability and is, thus, readily interpretable. Basing recommender trustworthiness estimation on the *FET* provides an additional advantage over the state-of-the-art: By substituting the original ver-

sion of *Fisher's Exact Test* for the *Fisher-Freeman-Halton Test* [58] gives an immediate extension of the method to the multinomial case.

The *FET*, as a measure for the similarity of two samples, will be leveraged in the following sections for two other purposes: It will be used in the conflict-aware fusion operation for computing the degree of conflict between two opinions (Definition 32, p. 118), and it is crucial in detecting changes in trustee behaviour when the assumption of stationary behaviour is forgone (Section 4.5, p. 160).

4.3 COMBINING AND AGGREGATING TRUSTWORTHINESS INFORMATION

In order to effectively use direct experience of a truster A and recommendations from recommenders, the knowledge contained in both has to be combined. In trust propagation, this is achieved through the use of the *discounting* and *consensus* operations, as presented, for instance, in [173]. In the following these two operations will be extended for application in the *Multinomial CertainTrust* model. Discounting and consensus are combined to evaluate the comparative performance of the various recommender trustworthiness approaches discussed above – the results of the evaluation can be found in Section 4.3.6, p. 126.

In addition, the *fusion* operation, used for aggregating opinions for which independence cannot be established, is extended in Section 4.3.3, p. 114. In particular, it is extended for use in *Multinomial CertainTrust*.

4.3.1 *Discounting*

Discounting represents weighting by scalar multiplication. The basic mechanism is applicable to binomial and multinomial samples and the corresponding sufficient statistics in the same manner as presented in [173]. Therefore, the operation remains fundamentally unchanged with the extension of *CertainTrust* to the multinomial model, which is presented in this thesis.

Definition 28 (Discounting)**.** Let $\delta \in [0;1]$ be a *discounting factor* and $\tilde{X} = (\vec{x}_1, \vec{x}_2, \vec{x}_3, \ldots, \vec{x}_{n-1}, \vec{x}_n)$ be a sample in $0-1$ *random vector* form with dimension $m \times n$. The *discounted* sample $\tilde{X}_{n,\delta}$ is computed as:

$$\tilde{X}_{n,\delta} = \delta \cdot \tilde{X} = (\delta \cdot \vec{x}_1, \delta \cdot \vec{x}_2, \delta \cdot \vec{x}_3, \ldots, \delta \cdot \vec{x}_{n-1}, \delta \cdot \vec{x}_n)$$

For the $m \in \mathbb{N}, m \geqslant 2$ row sums in $\tilde{X}_{n,\delta}$ it follows that for $1 \leqslant i \leqslant m$:

$$\tilde{\alpha}_i = \sum_{j=1}^{n} \delta \cdot x_{i,j} = \delta \cdot \sum_{j=1}^{n} x_{i,j} = \delta \cdot \alpha_i$$

with α_i – the row sum over the i-th row of $\tilde{X}$ – being a sufficient statistic for the i-th category in a multinomial proportion estimation problem (see, Section 3.2.2, p. 79).

Obviously, the resulting estimator for a multinomial proportion remains unchanged:

$$\hat{p}_i = \frac{\delta \cdot \alpha_i}{\delta \cdot \sum_{j=1}^{m} \alpha_j} = \frac{\alpha_i}{\sum_{j=1}^{m} \alpha_j}$$

However, the for the certainty estimation it appears as though the overall number of observations, n, has decreased[9], as:

$$\delta \cdot \sum_{i=1}^{m} \sum_{j=1}^{n} x_{i,j} = \delta \cdot n \leqslant n = \sum_{i=1}^{m} \sum_{j=1}^{n} x_{i,j}$$

Thus, the resulting discounted sample is given less weight, in the form of a lower certainty, than the original sample. This is so because the various certainty estimators introduced in this thesis (e.g., Section 3.2.3, p. 81), as well as those in the literature [108, 173, 197], scale in the number of observations. In combination with the consensus operation (Section 4.3.2, p. 113), discounting allows for the combination of observation conditional on their provenance.

Which, of course, leads to the question of how to actually determine the discounting factor δ. When the sample to be discounted comes from a recommender, as is usually the case in trust propagation, the δ should be a representation of the trustworthiness of the recommender. This trustworthiness is estimated by observing past performance of the recommender in the given context, just as in any other trustworthiness estimation task outlined before. However, the exact nature of *what exactly* is evaluated when establishing recommender trustworthiness deserves closer attention. This has been addressed in Section 4.2.

4.3.2 *Consensus*

In combination with discounting, the consensus operation provides the means for the aggregation of different opinions. In the binomial case, it compounds the sufficient statistics *sums of successes* and *sums of failures* into a single, new opinion. In the multinomial case, the sufficient statistics are given by the sums over the individual $m \geqslant 2, m \in \mathbb{N}$ categories. In the following, let $o_P^{R_i}$ denote an opinion of an entity R_i on another entity P. In the binomial case, the opinion is reporting the sufficient statistics r (*sum of successes*) and s (*sum of failures*) that R_i has experienced in interactions with P, $o_P^{R_i} = (r_P^{R_i}, s_P^{R_i})^{rs}$. In the multinomial case, the corresponding opinion is $o_P^{R_i} = ((\alpha_1)_P^{R_i}, (\alpha_2)_P^{R_i}, \ldots, (\alpha_m)_P^{R_i})^{\alpha}$.

Definition 29 (Consensus)**.** The consensus of opinions $o_P^{R_1}, o_P^{R_2}, \ldots, o_P^{R_n}$ for the general multinomial case is defined as

$$\begin{aligned} \text{consensus}(o_P^{R_1}, o_P^{R_2}, \ldots, o_P^{R_n}) &= \sum_{i=1}^{n} o_P^{R_i} = \\ &= o_P^{R_1} \oplus o_P^{R_2} \oplus \ldots \oplus o_P^{R_n} = \\ &= \left(\sum_{i=1}^{n} (\alpha_1)_P^{R_i}, \sum_{i=1}^{n} (\alpha_2)_P^{R_i}, \ldots, \sum_{i=1}^{n} (\alpha_m)_P^{R_i} \right) \end{aligned}$$

9 Recall, that $\tilde{X}$ is given in $0-1$ random vector form (Section 3.2.1, p. 80).

For the special case of binomial opinions, the consensus operator yields

$$consensus(o_P^{B_1}, o_P^{B_2}, \ldots, o_P^{B_n}) = \left(\sum_{i=1}^{n} r_P^{B_i}, \sum_{i=1}^{n} s_P^{B_i} \right)$$

The combination of consensus and discounting operators permits the discriminate aggregation of several different opinions into a new composite opinion. Let o_P^A and $o_P^{R_1}, o_P^{R_2}, \ldots, o_P^{R_n}$ $n+1$ different opinions, each with its corresponding discounting factor $\delta \in [0;1]$, so that δ_0 is the discounting factor[10] for opinion o_P^A and δ_i the discounting factor for opinion $o_P^{R_i}$. Then, the consensus operator yields

$$\begin{aligned} consensus(o_P^A, o_P^{R_1}, \ldots, o_P^{R_n}) &= \delta_0 \cdot o_P^A \oplus \sum_{i=1}^{n} \delta_i \cdot o_P^{R_i} = \\ &= \delta_0 \cdot o_P^A \oplus \delta_1 \cdot o_P^{R_1} \oplus \ldots \oplus \delta_n \cdot o_P^{R_n} = \\ &= \left(\delta_0 \cdot (\alpha_1)_P^A + \sum_{i=1}^{n} \delta_i \cdot (\alpha_1)_P^{R_i}, \ldots, \delta_0 \cdot (\alpha_n)_P^A + \sum_{i=1}^{n} \delta_i \cdot (\alpha_n)_P^{R_i} \right) \end{aligned}$$

The binomial consensus operator, in its basic form as presented above, has been extended by Ries [173] to increase its robustness against Sybil attacks. These improvements are applicable to the presented multinomial form without alterations and are given in their multinomial generalisations in Appendix D, p. 239.

Consensus and discounting enable the integration of direct experiences, that is, the observations made by truster A itself, and recommendations given by a set of recommenders, $R_1, \ldots, R_n$ on one specific trustee, P_j. The consensus operation provides the aggregation functionality, while discounting makes this aggregation conditional on the quality, in terms of trustworthiness, of the different recommenders. How to effectively compute the trustworthiness of recommenders is addressed in Section 4.2 and will be used to compare the various recommender trustworthiness estimation methods in Section 4.3.6.

4.3.3 *Fusion*

Beyond the consensus operator that aggregates opinions that are assumed to be independent into a new opinion through what essentially is a summation of observations, the need for an aggregation operation that allows for aggregating dependent opinions has been put forward. Jøsang's *Subjective Logic* [104, 105] introduces an operation for the *consensus of dependent opinions* that achieves aggregating dependent opinions through a certainty-dependent averaging operation. Following Jøsang, Ries et al.'s *CertainLogic* [175] adopted the

10 Normally, the truster's own opinion o_P^A is not discounted, i.e., $delta_0 = 1$. The parameter is given for the sake of completeness.

nomenclature for this operation by referring to it as *averaging fusion*, or simply *fusion*, in order to avoid confusion with the consensus operation that assumes independent opinions (as presented in Definition 29). The *CertainLogic* fusion operation presented in [175] is a straightforward adaptation of the *Subjective Logic* operation to the binomial *CertainTrust* opinion representation.

In the following, the fusion operation of *CertainLogic* will be considered from the perspective of observations and extended in two ways: firstly, a definition for the multinomial generalisation of the *CertainLogic* fusion operation[11] is given; secondly, the multinomial fusion operation is extended to handle preferences (in the form of weights), as well as potentially conflicting, contradictory opinions. The foundations of this extension have been published for the binomial fusion operation in [77]. Partly because the certainty estimators introduced in Chapter 3 are considerably more complex than those introduced in *Subjective Logic* or *CertainLogic*, the fusion operation will be leveraging the mapping from *CertainLogic* opinions to the *evidence space*, i.e., opinions of the form $o = (\alpha_1, \ldots, \alpha_m)^\alpha$, that is enabled by the bijection quality of the various certainty estimators.

One useful application of averaging fusion in the context of trust propagation is determining the average trustworthiness of a recommender. Because the fusion operation is essentially an arithmetic mean, one way to compute the overall opinion of A on recommender R_j as the fusion $o^A_{R_i} = fusion(o^A_{(R_i,P_1)}, \ldots, o^A_{(R_i,P_m)})$. This brings a limited degree of generalisability to estimating the trustworthiness of a recommender in situations where no prior recommendations on a specific trustee have been given by a recommender. In this case, the average trustworthiness of that recommender, computed from recommendations on other trustees, can serve as an informative prior.

In the remainder of this section (Section 4.3.3), let

$$o_1 = (\alpha^1_1, \ldots, \alpha^1_m)^\alpha; \ldots; o_n = (\alpha^n_1, \ldots, \alpha^n_m)^\alpha$$

be $n \in \mathbb{N}$ different, m-dimensional ($m \in \mathbb{N}, m \geqslant 2$) multinomial opinions, where α^i_j is the sum of observations that fall into category j for

11 Note that *Subjective Logic* already provides an equivalent multinomial representation.

opinion i. Let $\omega_1^m, \omega_2^m, \ldots, \omega_n^m$ be their corresponding *Multinomial CertainTrust* opinions, so that $\omega_k^m \equiv o_k$, with:

$$\omega_k^m = \left(\left(t_1^k = \frac{\alpha_1^k}{\sum_{j=1}^m \alpha_j^k}, c_1^k = c\left(x = \alpha_1^k, n = \sum_{j=1}^m \alpha_j^k\right)\right)_1; \right.$$
$$\left(t_2^k = \frac{\alpha_2^k}{\sum_{j=1}^m \alpha_j^k}, c_2^k = c\left(x = \alpha_2^k, n = \sum_{j=1}^m \alpha_j^k\right)\right)_2;$$
$$\ldots$$
$$\left. \left(t_m^k = \frac{\alpha_m^k}{\sum_{j=1}^m \alpha_j^k}, c_m^k = c\left(x = \alpha_m^k, n = \sum_{j=1}^m \alpha_j^k\right)\right)_m \right)$$

Then, the average fusion operation is defined as in the following Definition 30, building on [104].

Definition 30 (Average Fusion)**.** The fusion of opinions $o_1 = (\alpha_1^1, \ldots, \alpha_m^1); \ldots; o_n = (\alpha_1^n, \ldots, \alpha_m^n)$ for the general multinomial case is defined as

$$\begin{aligned} \text{fusion}(o^1, o^2, \ldots, o^n) &= \left(\frac{\sum_{i=1}^n \alpha_1^i}{n}, \frac{\sum_{i=1}^n \alpha_2^i}{n}, \ldots, \frac{\sum_{i=1}^n \alpha_m^i}{n} \right)^\alpha \\ &= \left(\alpha_1^{fusion}, \alpha_2^{fusion}, \ldots, \alpha_m^{fusion}\right)^\alpha \end{aligned}$$

Then, the fused *CertainTrust* opinion is given as

$$\omega_{fusion}^m = \left((t_1^{fusion}, c_1^{fusion})_1; \ldots; (t_m^{fusion}, c_m^{fusion})_m\right)$$

with

$$t_i^{fusion} = \frac{\alpha_i^{fusion}}{\sum_{j=1}^m \alpha_j^{fusion}}$$

and

$$\vec{c}^{fusion} = C^m(\alpha_1^{fusion}, \alpha_2^{fusion}, \ldots, \alpha_m^{fusion}; \sum_{j=1}^m \alpha_j^{fusion})$$

is the vector of certainty estimates computed according to one of the certainty estimators introduced in Section 3.2.3, p. 81, so that c_i^{fusion} equals the i-th component of vector $\vec{c}^{fusion}$

$$c_i^{fusion} = \vec{c}^{fusion}[i]$$

In Definition 30, the initial trust value parameters f have not been considered. This was done in order to provide a more natural mapping between the two opinion representations, that is, the relation $\omega_k^m \equiv o_k$. However, the fusion operation on initial trust value parameters $f_1^k, \ldots, f_m^k$ for *Multinomial CertainTrust* opinions of the form

$\omega_k^m = \left((t_1^k, c_1^k, f_1^k)_1; \ldots; (t_m^k, c_m^k, f_m^k)_m\right)$ is given as the simple average over each $f_1^k, \ldots, f_m^k$, for all k; that is:

$$f_i^{fusion} = \frac{\sum_{j=1}^{n} f_i^j}{n}$$

Therefore, the resulting fused opinion ω_{fusion}^m has the following complete form:

$$\omega_{fusion}^m = \left((t_1^{fusion}, c_1^{fusion}, f_1^{fusion})_1; \ldots; (t_m^{fusion}, c_m^{fusion}, f_m^{fusion})_m\right)$$

In order to express preference of one opinion over another, or an ordering of preferences over multiple opinions, the average fusion operation can be adapted into a *weighted average fusion* variant. For this, assume a multiplicative factor w_i for each opinion $\omega_i^m \equiv o^i = (\alpha_1^i, \ldots, \alpha_m^i)$. Recalling that the different α_j^i are sufficient statistics of a sample of multinomial observations, the result of multiplying an opinion o^i with a scalar value w_i is equivalent in effect to the discounting operation described in Definition 28, p. 112; the single difference being that the discounting factor $\delta_i \in [0; 1]$, while $w_i \in \{0, \mathbb{R}^+\}$. Consequently, the *weighted average fusion* can be defined as:

Definition 31 (Weighted Average Fusion)**.** The weighted fusion of opinions $o_1 = (\alpha_1^1, \ldots, \alpha_m^1); \ldots; o_n = (\alpha_1^n, \ldots, \alpha_m^n)$ with weights $w_1, \ldots, w_n, w_i \in \{0, \mathbb{R}^+\}, \sum_{i=1}^{n} w_i \neq 0$ for the general multinomial case is defined as

$$\begin{aligned} &w.\text{fusion}(o^1, o^2, \ldots, o^n; w_1, \ldots, w_n) = \\ &\quad = \left(\frac{\sum_{i=1}^{n} w_i \cdot \alpha_1^i}{\sum_{i=1}^{n} w_i}, \ldots, \frac{\sum_{i=1}^{n} w_i \cdot \alpha_m^i}{\sum_{i=1}^{n} w_i}\right)^{\alpha} = \\ &\quad = \left(\alpha_1^{w.\text{fusion}}, \ldots, \alpha_m^{w.\text{fusion}}\right)^{\alpha} \end{aligned}$$

Then, the fused *CertainTrust* opinion under weighted fusion is given as

$$\omega_{w.\text{fusion}}^m = \left((t_1^{w.\text{fusion}}, c_1^{w.\text{fusion}})_1; \ldots; (t_m^{w.\text{fusion}}, c_m^{w.\text{fusion}})_m\right)$$

with

$$t_i^{w.\text{fusion}} = \frac{\alpha_i^{w.\text{fusion}}}{\sum_{j=1}^{m} \alpha_j^{w.\text{fusion}}}$$

and

$$\vec{c}^{w.\text{fusion}} = C^m(\alpha_1^{w.\text{fusion}}, \alpha_2^{w.\text{fusion}}, \ldots, \alpha_m^{w.\text{fusion}}; \sum_{j=1}^{m} \alpha_j^{w.\text{fusion}})$$

is the vector of certainty estimates computed according to one of the certainty estimators introduced in Section 3.2.3, p. 81, so that $c_i^{w.\text{fusion}}$ equals the i-th component of vector $\vec{c}^{w.\text{fusion}}$

$$c_i^{w.\text{fusion}} = \vec{c}^{w.\text{fusion}}[i]$$

The weighted fusion operation on the initial trust values f is, analogously to the un-weighted average fusion operation, an average over the individual f_i^j. To account for the weights, the regular average is replaced by a weighted average, yielding

$$f_i^{w.fusion} = \frac{\sum_{j=1}^{n} w_i \cdot f_i^j}{\sum_{j=1}^{n} w_i}$$

Weighting permits expressing preferences for particular opinions over others during the process of fusion. While this is useful extension to the average fusion operator because it enables an entity to actively influence the fusion by incorporating external information into the weights, it only represents an intermediate step. In order to account for dissimilarities, or *conflict*, between the opinions that are to be fused, [77] introduced a *conflict-aware* extension to the fusion operation.

The measure of conflict, DoC, for this binomial *conflict-aware fusion operation* was defined based on the average residuals between all combinations of $n \in \mathbb{N}$ binomial opinions $\omega_1 = (t^1, c^1), \ldots, \omega_n = (t^n, c^n)$ with weights $w_1, \ldots, w_n$:

$$\mathrm{DoC} = \frac{\sum_{i=1,j=1} \mathrm{DoC}_{\omega_i,\omega_j}}{\frac{n\cdot(n-1)}{2}} \tag{12}$$

where

$$\mathrm{DoC}_{\omega_i,\omega_j} = \left|t^i - t^j\right| \cdot c^i \cdot c^j \cdot \left(1 - \left|\frac{w_i - w_j}{w_i + w_j}\right|\right) \tag{13}$$

Formulating a measure of conflict for the multinomial case in a similar manner may be achieved by leveraging $\mathcal{L}_1$ vector norms. However, this would still involve the explicit pairwise computation of the DoC. In order to facilitate the multinomial extension of the conflict-aware weighted fusion operation in a more compact way, another (dis-)similarity measure is applied in the following – the p-value returned by the extended Fisher's Exact Test [58] (or *Fisher-Freeman-Halton Test*).

As already described in Definition 26, p. 108, the (extended) Fisher's Exact Test (FET) returns as its p-value the probability that several m-categorical multinomial samples are independent, i.e., that these samples were generated by the same categorical process. A high p-value thus indicates a high similarity of the samples, while a low p-value is indicative of a high degree of dissimilarity, or conflict. Using the extended FET as a means for computing the degree of conflict, DoC, leads to the following definition of a multinomial conflict-aware weighted fusion operation:

Definition 32 (Conflict-Aware Weighted Average Fusion)**.** The conflict-aware weighted fusion of opinions $o_1 = (\alpha_1^1, \ldots, \alpha_m^1); \ldots; o_n = (\alpha_1^n, \ldots, \alpha_m^n)$

with weights $w_1, \ldots, w_n, w_i \in \{0, \mathbb{R}^+\}, \sum_{i=1}^{n} w_i \neq 0$ for the general multinomial case is defined as

$$\begin{aligned} &c.fusion(o^1, \ldots, o^n; w_1, \ldots, w_n) = \\ &= \left(f(\mathfrak{p}) \cdot \frac{\sum_{i=1}^{n} w_i \cdot \alpha_1^i}{\sum_{i=1}^{n} w_i}, \ldots, f(\mathfrak{p}) \cdot \frac{\sum_{i=1}^{n} w_i \cdot \alpha_m^i}{\sum_{i=1}^{n} w_i} \right)^{\alpha} = \\ &= \left(f(\mathfrak{p}) \cdot \alpha_1^{w.fusion}, \ldots, f(\mathfrak{p}) \cdot \alpha_m^{w.fusion} \right)^{\alpha} = \\ &= \left(\alpha_1^{c.fusion}, \ldots, \alpha_m^{c.fusion} \right)^{\alpha} \end{aligned}$$

where $1 - \mathfrak{p}$ is the p-value returned by the extended FET (Fisher-Freeman-Halton Test) of independence for opinions $o_1, \ldots, o_n$, and $f(\mathfrak{p})$ is a function of $\mathfrak{p}$ that satisfies $f(\mathfrak{p}) \in [0; 1]$. In case that the opinions $o_1, \ldots, o_n$ are assigned weights $w_1, \ldots, w_n$, the extended FET is applied to the weighted opinions $o_1 = (w_1 \cdot \alpha_1^1, \ldots, w_1 \cdot \alpha_m^1); \ldots; o_n = (w_n \cdot \alpha_1^n, \ldots, w_n \cdot \alpha_m^n)$.

Then, the fused *CertainTrust* opinion under conflict aware weighted fusion is given as

$$\omega_{c.fusion}^m = \left((t_1^{c.fusion}, c_1^{c.fusion})_1; \ldots; (t_m^{c.fusion}, c_m^{c.fusion})_m \right)$$

with

$$t_i^{c.fusion} = \begin{cases} 0.5 & \text{if } f(\mathfrak{p}) = 0 \\ \frac{\alpha_i^{w.fusion}}{\sum_{j=1}^{m} \alpha_j^{w.fusion}} = \frac{\alpha_i^{c.fusion}}{\sum_{j=1}^{m} \alpha_j^{c.fusion}} & \text{else} \end{cases}$$

and

$$\vec{c}^{c.fusion} = C^m \left(\alpha_1^{c.fusion}, \alpha_2^{c.fusion}, \ldots, \alpha_m^{c.fusion}; \sum_{j=1}^{m} \alpha_j^{c.fusion} \right)$$

is the vector of certainty estimates computed according to one of the certainty estimators introduced in Section 3.2.3, p. 81, so that $c_i^{c.fusion}$ equals the i-th component of vector $\vec{c}^{c.fusion}$

$$c_i^{c.fusion} = \vec{c}^{c.fusion}[i]$$

Furthermore, for the initial trust parameters it holds that:

$$f_i^{c.fusion} = f_i^{w.fusion}$$

EXAMPLE The degree of conflict is introduced to the fusion operation through a multiplicative factor, determined by $f(\mathfrak{p}) \in [0; 1]$. The value of $f(\mathfrak{p})$ represents a measure of the similarity of the the individual opinions $o^1, \ldots, o^n$, so that the more similar the opinions, the more the value of $f(\mathfrak{p})$ approaches one. In the most basic case we consider, $f(\mathfrak{p}) = id(\mathfrak{p}) = \mathfrak{p}$ equals 1 minus the p-value returned by the Fisher-Freeman-Halton Test of independence. This, as has been outlined above in Section 4.2.4, p. 106, is the statistical probability that opinions $o^1, \ldots, o^n$ were generated by the same random process.

The procedure of introducing the degree of conflict as a multiplicative factor can be motivated by the following example. Suppose a truster A receives a recommendation from recommender R_i on some potential trustee P_j. Furthermore, assume that R_i has supplied recommendations on a number of *other* trustees, $P_1, \ldots, P_n$, but not on P_j. Therefore, A has $n \in \mathbb{N}$ different opinions on the trustworthiness of recommender R_i in providing recommendations on the individual trustees $P_1, \ldots, P_n$, $\omega^A_{(R_i,P_1)}, \ldots, \omega^A_{(R_i,P_n)}$. However, A has no opinion on R_i's ability to provide reliable recommendations on P_j. Applying a fusion operation on opinions $o^A_{(R_i,P_1)}, \ldots, o^A_{(R_i,P_n)}$, A can generalise from R_i's trustworthiness with regard to other trustees in order to derive an opinion on R_i's ability to recommend P_j.

Now assume that exactly half of the opinions $\omega^A_{(R_i,P_1)}, \ldots, \omega^A_{(R_i,P_n)}$ take the form $(1, 1)$, that is, the trust estimate t of these opinions equals 1 at a certainty value of 1, while the other half take the form $(0, 1)$. In other words, recommender R_i is very good at recommending for half of the population of trustees, while being very bad at recommending the other half of trustees. Additionally, truster A is highly certain about about these trustworthiness estimates about R_i, as it has received numerous recommendations from R_i and has considerable direct experience with interaction partners P_1 through P_n. Applying averaging fusion to $\omega^A_{(R_i,P_1)}, \ldots, \omega^A_{(R_i,P_n)}$, as per Definition 30, p. 116, yields a fused opinion $\omega = (0.5, 1)$. This would indicate that R_i's overall trustworthiness in recommending is $t = 0.5$ and that this estimate can be made with high certainty. However, for purposes of predicting the performance of R_i in recommending a specific trustee, such as P_j, the high certainty value can be undesirable. Looking at the evidence of R_i's quality in recommending trustees, it is highly likely that its recommendations on P_j will be either spot on or entirely off. Since we do not consider any external discriminators to correlate whether R_i will provide a very good or very bad recommendation on P_j, however, either alternative cannot be predicted accurately or attributed with a high certainty value.

Using the conflict-aware fusion operation of Definition 32, p. 118 accounts for this by decreasing the certainty of the fused opinion by multiplication with the degree of conflict – that is, the conflict-aware fusion operation reduces the confidence in the accuracy of the estimate t. Assuming two equally weighted, contradicting opinions, that is, two opinions with a high degree of conflict, the certainty in the resulting fused opinion will be decreased proportional to the degree of conflict.

4.3.4 *Evaluation: Comparison of degree of conflict computation in the binomial case*

The following Tables 5 and 6 show the degree of conflict computed for two binary opinions, $o_1 = (r_1, s_1)^{rs}$ and $o_2 = (r_2, s_2)^{rs}$. In Table 5, the total number of experiences per opinion is $n = r_1 + s_1 = r_2 + s_2 = 10$, in Table 6 $n = r_1 + s_1 = r_2 + s_2 = 100$. The tables' rows vary the proportion of r_1 to s_1, while the columns vary the proportion of r_2 to s_2. The fields of the tables give the degree of conflict: first as computed per Equation 13 , p. 118 (labelled 'CT'), and then based on the p-value of the *Fisher-Freeman-Halton Test / FET*, or *Fisher Score* (labelled 'FS'). It can be seen that the *FET*-based method is considerably more conservative in that it generally yields a higher degree of conflict than the method of Equation 13.

The *FET*-based degree of conflict (Fisher Score) takes the size of the opinions into account. This can be seen by comparing the values computed for the same proportions given different opinion sizes. Table 5, p. 122, and Table 6, p. 123, illustrate this. Comparing the degrees of conflict for the same proportions $t = 0.1, t = 0.2, \ldots, t = 1$ between the two tables, i.e., for two different total numbers of observations, $n = 10$ and $n = 100$, it is obvious that the statistical properties of the Fisher Score and the degree of conflict computation in Equation 13 differ. Recalling that the *FET* computes the exact probability that the two opinions were generated from the same random process, this reflects the increase in the certainty of the estimates as the number of observations increases[12].

Generally, the *FET*-based degree of conflict reports considerably higher degree of conflict values than the method in Equation 13. in Figure 6, the degree of conflict (depicted on the vertical axis) is plotted against the percentage of successes in each of the opinions o_1 and o_2, i.e., $\frac{r_1}{n} \cdot 100\%$ and $\frac{r_2}{n} \cdot 100\%$ for $n = 100$, on the horizontal axes. Figure 6a shows a much slower increase in the degree of conflict as $\frac{r_1}{n}$ diverges from $\frac{r_2}{n}$ than does Figure 6b. That is, the FET-based Fisher Score is considerably more conservative than the methods used in Equation 13.

That is, the probability that the test reports for the independence of two opinions is normally lower than the ad-hoc score returned by the method in Equation 13 [77]. In order to overcome this conservativeness – that may be considered excessive for real-world applications – the Fisher Score similarity measure, denoted $f(p)$, can be modified accordingly. As mentioned previously (Definition 32, p. 118), the most basic version of the function to compute a degree of conflict based on the *FET* is $f(p) = id(p) = p$, where $1 - p$ is the p-value returned by the *FET* for independence. In order to mimic the more liberal behaviour

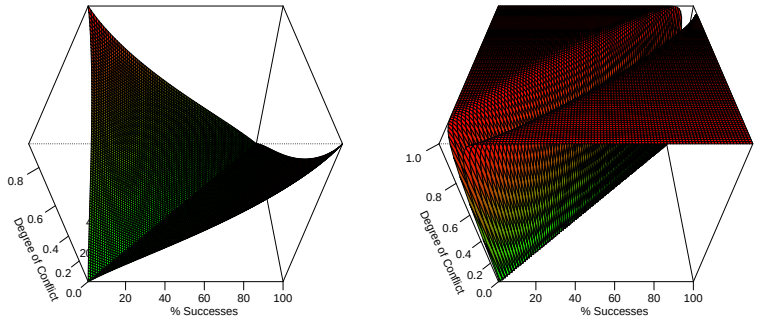
12 Compare also the monotonicity property of the certainty estimators introduced in Chapter 3.

	r_2 =1	r_2 =2	r_2 =3	r_2 =4	r_2 =5	r_2 =6	r_2 =7	r_2 =8	r_2 =9	r_2 =10
	CT | FS	CT | FS	CT | FS	CT | FS	CT | FS	CT | FS	CT | FS	CT | FS	CT | FS	CT | FS
r_1 =1	0 | 0	0.03 | 0	0.06 | 0.42	0.09 | 0.7	0.12 | 0.86	0.15 | 0.94	0.19 | 0.98	0.23 | 0.99	0.3 | 1	0.4 | 1
r_1 =2	0.03 | 0	0 | 0	0.03 | 0	0.05 | 0.37	0.08 | 0.65	0.11 | 0.83	0.14 | 0.93	0.18 | 0.98	0.23 | 0.99	0.32 | 1
r_1 =3	0.06 | 0.42	0.03 | 0	0 | 0	0.02 | 0	0.05 | 0.35	0.07 | 0.63	0.1 | 0.82	0.14 | 0.93	0.19 | 0.98	0.26 | 1
r_1 =4	0.09 | 0.7	0.05 | 0.37	0.02 | 0	0 | 0	0.02 | 0	0.05 | 0.34	0.07 | 0.63	0.11 | 0.83	0.15 | 0.94	0.21 | 0.99
r_1 =5	0.12 | 0.86	0.08 | 0.65	0.05 | 0.35	0.02 | 0	0 | 0	0.02 | 0	0.05 | 0.35	0.08 | 0.65	0.12 | 0.86	0.17 | 0.97
r_1 =6	0.15 | 0.94	0.11 | 0.83	0.07 | 0.63	0.05 | 0.34	0.02 | 0	0 | 0	0.02 | 0	0.05 | 0.37	0.09 | 0.7	0.14 | 0.91
r_1 =7	0.19 | 0.98	0.14 | 0.93	0.1 | 0.82	0.07 | 0.63	0.05 | 0.35	0.02 | 0	0 | 0	0.03 | 0	0.06 | 0.42	0.11 | 0.79
r_1 =8	0.23 | 0.99	0.18 | 0.98	0.14 | 0.93	0.11 | 0.83	0.08 | 0.65	0.05 | 0.37	0.03 | 0	0 | 0	0.03 | 0	0.08 | 0.53
r_1 =9	0.3 | 1	0.23 | 0.99	0.19 | 0.98	0.15 | 0.94	0.12 | 0.86	0.09 | 0.7	0.06 | 0.42	0.03 | 0	0 | 0	0.04 | 0
r_1 =10	0.4 | 1	0.32 | 1	0.26 | 1	0.21 | 0.99	0.17 | 0.97	0.14 | 0.91	0.11 | 0.79	0.08 | 0.53	0.04 | 0	0 | 0

Table 5: Degree of Conflict according to [77] (labelled 'CT') and Fisher Score (labelled 'FS'), n=10

	r_2 =10	r_2 =20	r_2 =30	r_2 =40	r_2 =50	r_2 =60	r_2 =70	r_2 =80	r_2 =90	r_2 =100
	CT | FS	CT | FS	CT | FS	CT | FS	CT | FS	CT | FS	CT | FS	CT | FS	CT | FS	CT | FS
r_1 =10	0 | 0	0.07 | 0.93	0.15 | 1	0.21 | 1	0.28 | 1	0.36 | 1	0.44 | 1	0.52 | 1	0.62 | 1	0.76 | 1
r_1 =20	0.07 | 0.93	0 | 0	0.07 | 0.86	0.14 | 1	0.2 | 1	0.27 | 1	0.35 | 1	0.43 | 1	0.52 | 1	0.65 | 1
r_1 =30	0.15 | 1	0.07 | 0.86	0 | 0	0.07 | 0.82	0.13 | 0.99	0.2 | 1	0.27 | 1	0.35 | 1	0.44 | 1	0.55 | 1
r_1 =40	0.21 | 1	0.14 | 1	0.07 | 0.82	0 | 0	0.07 | 0.8	0.13 | 0.99	0.2 | 1	0.27 | 1	0.36 | 1	0.47 | 1
r_1 =50	0.28 | 1	0.2 | 1	0.13 | 0.99	0.07 | 0.8	0 | 0	0.07 | 0.8	0.13 | 0.99	0.2 | 1	0.28 | 1	0.39 | 1
r_1 =60	0.36 | 1	0.27 | 1	0.2 | 1	0.13 | 0.99	0.07 | 0.8	0 | 0	0.07 | 0.82	0.14 | 1	0.21 | 1	0.31 | 1
r_1 =70	0.44 | 1	0.35 | 1	0.27 | 1	0.2 | 1	0.13 | 0.99	0.07 | 0.82	0 | 0	0.07 | 0.86	0.15 | 1	0.24 | 1
r_1 =80	0.52 | 1	0.43 | 1	0.35 | 1	0.27 | 1	0.2 | 1	0.14 | 1	0.07 | 0.86	0 | 0	0.07 | 0.93	0.16 | 1
r_1 =90	0.62 | 1	0.52 | 1	0.44 | 1	0.36 | 1	0.28 | 1	0.21 | 1	0.15 | 1	0.07 | 0.93	0 | 0	0.08 | 1
r_1 =100	0.76 | 1	0.65 | 1	0.55 | 1	0.47 | 1	0.39 | 1	0.31 | 1	0.24 | 1	0.16 | 1	0.08 | 1	0 | 0

Table 6: Degree of Conflict according to [77] (labelled 'CT') and Fisher Score (labelled 'FS'), n=100



(a) Degree of conflict according to Equation 13 [77]

(b) FET-based degree of conflict

Figure 6: Comparison of degree of conflict computations, $n = 100$.

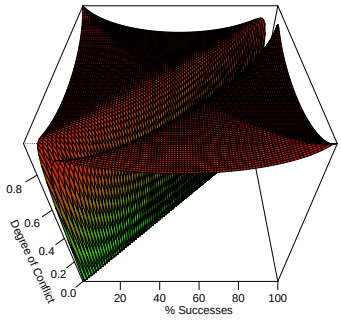
of the original *CertainTrust* degree of conflict computation [77], various functions of p could reasonably be applied.

In order to control the slope of the Fisher Score-based degree of conflict function, a multiplicative factor can be introduced to $f(p)$. One intuition that can be leveraged in order to determine the factor is the heuristic already used in [77]; that is, that the slope should be in some way proportional to the distance of $\frac{r_1}{n}$ from $\frac{r_2}{n}$ and to account for the certainty in each of the two opinions o_1 and o_2. Considering the computation of the degree of conflict in Equation 13, which contains as factors the product of the certainties, $c_1 \cdot c_2$, and the absolute difference of the trust scores, $|t_1 - t_2|$, an analogously designed function $f(p)$, for the degree of conflict of binomial opinions, can easily be constructed. Equation 14 provides an ad-hoc way of controlling the slope, and hence the conservativeness, of the *FET*-based degree of conflict function:

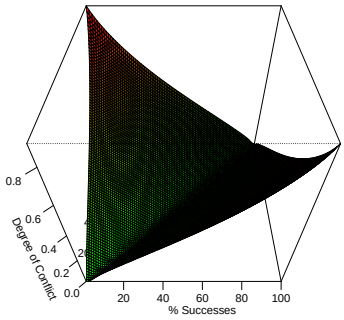
$$f(p) = p \cdot (c_1 \cdot c_2)^x \cdot (|t_1 - t_2|)^y \tag{14}$$

with $x, y \in \mathbb{R}_0^+$. Equation 14 provides a set of equations that can be produced by instantiating parameters x and y. Figure 7 shows the behaviour of the degree of conflict under different instantiations of Equation 14, as well as $f(p) = p \cdot min(c_1, c_2)$. As can be seen, the multiplicative factor has a considerable effect on the shape of the degree of conflict function. Whether or not this is desirable depends on the application. The unmodified Fisher Score represents the a conservative, and by its computation exact way to measure the independence of two opinions. The introduction of a multiplicative factor can alleviate the Fisher Score's conservativeness; the price for this, however, is that the measure becomes arbitrary.

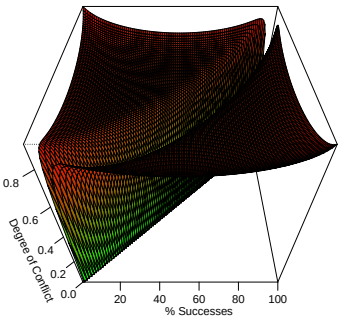
The more conservative nature of the *FET*-based degree of conflict computation has a direct impact on the fusion of opinions. This means that the uncertainty reported under conflict-aware fusion using an



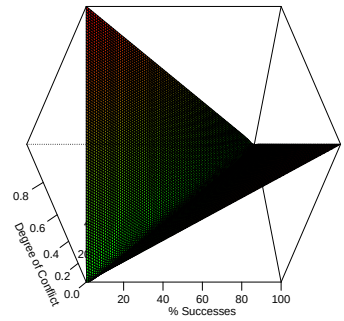
(a) $f(p) = p \cdot \min(c_1, c_2)$



(b) $f(p) = p \cdot (c_1 \cdot c_2) \cdot (|t_1 - t_2|)$



(c) $f(p) = p \cdot (c_1 \cdot c_2)$



(d) $f(p) = p \cdot (|t_1 - t_2|)$

Figure 7: Various instantiations of Equation 14 for $n = 100$.

unmodified, *FET*-based degree of conflict will report a considerably lower certainty score. As per Definition 32, this difference in certainty is proportional to the difference in the degree of conflict scores, for instance as reported in Tables 5, p. 122, and 6, p. 123. Obviously, as the degree of conflict is a multiplicative constant in Definition 32, the computation of the trust scores remains unaffected.

4.3.5 *Recommender Trustworthiness as a Discounting Factor*

The principal goal of determining recommender trustworthiness in trust propagation is its use a discounting factor in the consensus operation. If the estimate of the trustworthiness of a recommender is plugged *directly* into the consensus operation as a discounting factor, however, highly untrustworthy recommenders would still have an impact on the trustworthiness estimation. This is highly undesirable.

In fact, a trust score in the range $[0; 0.5[$ indicates active mistrust in the actions of the trustee, while a trust score of exactly 0.5 indicates a neutral attitude. Only in the range of $]0.5; 1]$ is there an active trusting attitude from truster to trustee. Therefore, discounting with the recommender trustworthiness directly would insinuate active trust, where there actually is a degree of active *mistrust*. In order to accommodate for this, and to increase the robustness of the consensus operation [173], the discounting factor δ_i attributed to R_i is defined as:

Definition 33 (Robust Discounting with Recommender Trustworthiness [173])**.** Let $t_e \in [0; 1]$ be a threshold that indicates the lower bound of active trust in recommendations (as a default, assume 0.5). Let $E(\omega^A_{R_i})$ be the *CertainTrust* expectation value of A, indicating the trustworthiness of recommender R_i when providing recommendations on P_j

Then the robust discounting factor δ_i attributed to R_i by A is defined as:

$$\delta_i = \begin{cases} 0 & \text{if } n \leqslant t_e \\ \frac{1}{1-t_e} \cdot (E(\omega^A_{R_i}) - t_e) & \text{else} \end{cases}$$

4.3.6 *Evaluation: Comparing Recommender Trustworthiness Estimation Approaches in Trust Propagation*

In order to establish the overall trustworthiness of a recommender, its performance over all recommendations within a specific context has to be considered. For this let $P = \{P_1, P_2, \ldots, P_m\}$ be the set of all potential interaction partners within context $C(I)$. An overall score of the trustworthiness of a specific recommender R_i is then computed by calculating the R_i's trustworthiness with regard to the individual recommendations on $P_1, P_2, \ldots, P_m$ and using a fusion operation (see

Section 4.3.3) to provide an average trust and certainty score on the ability of R_i to recommend correctly.

The basis for the averaged trust and certainty scores are the individual trustworthiness estimates computed for the recommendation performance of R_i on $P_1, P_2, \ldots, P_m$. In Section 4.2, various methods for computing these estimates were presented. Under assumptions of stationarity, the different methods of computing recommender trustworthiness (see also Section 4.2, pps. 100) are compared against a base truth in the following. To briefly recall, these methods are:

- *tendency*,
- *linear*,
- *max-certainty*,
- *sensitivity*,
- *average-β*, and
- *FET-based recommender trustworthiness*

For the evaluation, a time series of interactions between a truster A and a trustee P was simulated. The estimand parameter p, the true trustworthiness of trustee P, provides the base truth against which the trust estimates are compared. During each time step truster A interacts with trustee P, thus building its own interaction history of direct experiences with P, expressed as opinion o_P^A. Additionally a recommender R provides recommendations on P to A. These recommendations represent R's opinion on A, based on R's prior experience with P. Between two time steps, the recommendation by R will be updated with new experiences, the number of which is determined by a Poisson distributed random variable. A's final estimate of the trustworthiness of P is computed at each time step as the consensus of A's direct opinion on P with R's recommendation on P, which in turn is discounted using the robust discounting factor (Section 4.3.5) based on the various recommender trustworthiness measures (Section 4.2). That is, at each time step an opinion $o_P^{A \oplus R}$ is computed as:

$$o_P^{A \oplus R} = o_P^A \oplus \delta \cdot o_P^R$$

In a Monte-Carlo simulation, the performance of the recommender trustworthiness estimation methods from Section 4.2 was evaluated by comparing the value of the trust estimate $t = \hat{p}$ against the true parameter p. In the following, the predictive performance is reported in terms of the root mean squared error (*rmse*) of the estimate $t = \hat{p}$ compared to the true parameter p:

Definition 34. Let $\mathcal{P} = (p_1, p_2, \ldots, p_n)$ be a vector of observed values of the trustworthiness of trustee P and $\hat{\mathcal{P}} = (\hat{p}_1, \hat{p}_2, \ldots, \hat{p}_n)$ a vector

of corresponding trust estimates at time step $1, 2, \ldots, n$. Then the root mean squared error is defined as:

$$rmse = \sqrt{\frac{1}{n} \cdot \sum_{i=1}^{n} (\hat{p}_i - p_i)^2}$$

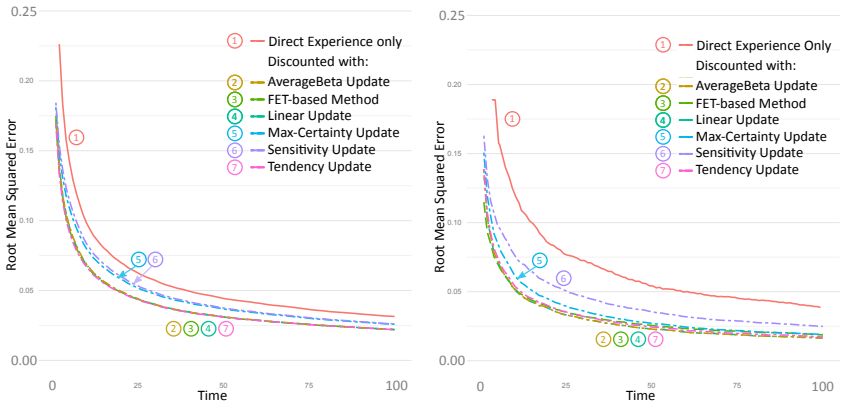
The predictive performance of A's own interaction history of direct experiences, in the form of opinion o_P^A, served as a comparative baseline in order to assess the usefulness of the recommender trustworthiness estimation approaches.

HONEST RECOMMENDER, $p = 0.5$ In the initial simulation, trustee P generates binomial feedback with a probability of success of $p = 0.5$ and recommender R behaves honestly, that is, the interactions it reports were generated at the same value[13] of $p = 0.5$. Figure 8, p. 129, depicts the smoothed outcome of a 10,000 run Monte-Carlo simulation with a Poisson distributed random variable for the number of additional recommendations per time step determined by Poisson parameter $\lambda = 1$ (Figure 8a). The results did not change qualitatively with larger λ-values (Figure 8b) or an increased number of runs. As can be seen from the figure, all methods provided a significant improvement over the baseline of only using direct experience information, that is, of using only o_P^A. However, the *max-certainty* and *sensitivity* performed significantly worse, according to a *Wilcoxon-Mann-Whitney* test [203] (one-sided significance value < 0.001), than the other methods. *Tendency*, *linear*, *average*-β and the *FET-based* recommender trustworthiness methods perform at similar performance levels and do not exhibit statistically significant differences. With an increasing number of evidences in the recommendations, that is, a higher factor λ, the *max-certainty* method showed improved performance, while the *sensitivity* update method did not.

HONEST RECOMMENDER, RANDOM p When randomising the true trustworthiness p of potential interaction partner P, so that each run in a 10,000 run Monte-Carlo Simulation is conducted with its own uniformly generated $p \in [0; 1]$, the qualitative performance of the different recommender trustworthiness estimation approaches did not differ considerably from the performance reported under constant p. As in the previous simulations, the recommender provided honest feedback. Figure 9, p. 129 shows the resulting *rmse*.

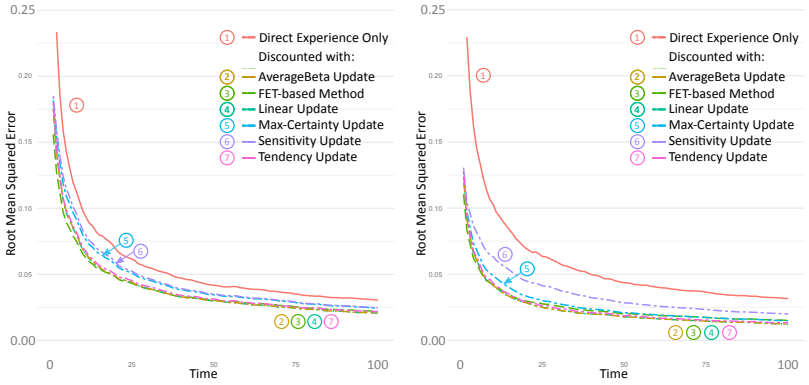
When assuming honesty on part of the recommender, all recommender trustworthiness estimation methods perform reasonably well. *Sensitivity* and *max-certainty* updates are slightly but significantly outperformed by the other methods; however, all methods provide sta-

13 Simulations were also conducted at other trustee trustworthiness values; the general relative performance of the recommender trustworthiness estimation approaches was unaffected.



(a) $p = 0.45, \lambda = 1$ (b) $p = 0.5, \lambda = 5$

Figure 8: Root Mean Squared Error of the prediction quality under various recommender trustworthiness estimation mechanisms, honest recommender.



(a) Random p, $\lambda = 1$ (b) Random p, $\lambda = 5$

Figure 9: Root Mean Squared Error of the prediction quality under various recommender trustworthiness estimation mechanisms, honest recommender.

tistically significant improvement over using only direct experience to gauge the quality of trustee P. Specifically, the estimation error converges to zero over time, for all estimation methods.

Yet, because the recommender was actually honest in the simulations, the presented simulations do not reveal the behaviour of the recommender trustworthiness estimation methods under malicious or accidentally misreporting recommenders. Rather, the preceding simulations only provide an insight into the way that the methods deal with the inherent sampling error. Therefore, the performance under (dishonest) misreporting on part of the recommender will be investigated in the following.

DISHONEST RECOMMENDER, $p = 0.5$ A misreporting or dishonest recommender reports a different estimated trustworthiness value p of trustee to truster A. In the simulation, this has been realised by having the misreporting or dishonest recommender generate binary feedback by executing a Bernoulli process at a fixed offset from the true trustee trustworthiness p. Thus, the feedback that a misreporting or dishonest recommender returns is generated according to a Binomial distribution, $\mathtt{Bin}(n, p + \mathtt{offset})$ for a recommendation consisting of $n \in \mathbb{N}$ reported experiences, with $p + \mathtt{offset} \in [0; 1]$.

Fixing the probability of success p, i.e., the true trustworthiness of trustee P, at 0.5, Figure 10, p. 132, and Figure 11, p. 133, show the performance[14] of the different recommender trustworthiness estimation methods under misreporting recommenders. The subfigures vary in the degree of offset the recommender exhibits, compared to the true value of p, and the number of experiences the recommender generates, based on a Poisson distribution with parameter λ, in each time step.

The simulation results in Figure 10 assume that the recommender misreport the trustee's performance from the start, that is, beginning at time step 1 it generates binary feedback at a probability of $p + \mathtt{offset}$. As is evident from the Figure 10, the performance of the different estimation methods is dependent both on the offset and the number of evidences that constitute the recommendation, i.e., the parameter λ. Under a small offset ($p + 0.05$) and a modest $\lambda = 1$, the performance of the estimation methods is not adversely affected (Figure 8a, p. 129). The slight offset is compensated for by a relatively equal balance of direct experience and recommendation. The performance impact of the offset is not distinguishable from general sampling error. Utilising recommendations still provides faster convergence of the *rmse* than using only direct experience, irrespective of the method chosen.

Increasing λ to 5, while maintaining an offset of 0.05 (Figure 10b), show first signs that the misreported probability of success impacts

14 Performance measured in terms of the *rmse*.

the estimation result for estimand p. While most recommender estimation approaches (*FET-based*, *max-certainty*, *sensitivity*, and *tendency*) still show convergent behaviour and the *rmse* is lower for all recommender estimation approaches than it is for the direct experience only estimation, *linear* and *average*-β begin to diverge.

Increasing the offset shows has a direct impact on the performance of using recommendations. With a moderate offset of $p + 0.1$ and $\lambda = 1$, *FET-based*, *max-certainty*, *sensitivity*, and *tendency* show decreased efficiency but maintain performance slightly better than just using direct experience. *Linear* and *average*-β start to exhibit performance inferior to the use direct experience only (Figure 10c). Increasing λ to 5, makes the performance differences more apparent, with all approaches losing their performance edge over direct experience only estimation (Figure 10d).

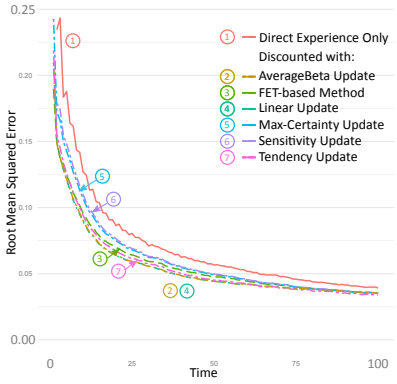
At a larger offset, $p + 0.25$, the effects are more pronounced. For $\lambda = 1$, the *FET-based*, *max-certainty*, and *sensitivity* methods perform at about the same level as the direct experience only approach, exhibiting only non-significant differences according to a Wilcoxon test. The other methods perform significantly worse (Figure 10e). For $\lambda = 5$, all methods perform significantly worse than the direct experience only method. Particularly the *linear*, *average*-β and *tenedency* update methods do not show any convergence behaviour at all (Figure 10f).

in Figure 11, the simulation scenario is varied. Here, the recommender gives honest recommendations from time step 0 to time step 50, after which it reports recommendations generated under an offset from the true value of p. The recommender thus initially establishes its trustworthiness before beginning to misreport. Qualitatively, the results mimic those of the previous simulation setup, with the *sensitivity* and *FET-based* methods providing the most robust performance. The *FET-based* provides a good mix of quick convergence during the first 50 time steps (during which the recommender reports honestly) and robustness during the latter 50 time steps (during which the recommender misreports).

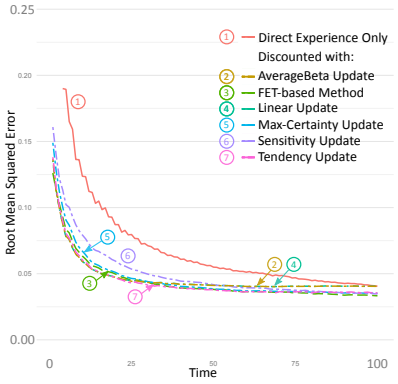
DISHONEST RECOMMENDER, RANDOM p Randomizing the trustee trustworthiness parameter p so that $p \in [0; 1]$ is a uniformly distributed random variable. As in the previous simulations setups, the simulation was repeated 10,000 times with a random but fixed p for each repeat. The misreporting recommender generates recommendations from a Bernoulli process with a probability of success of

$$max(min(p + offset, 1), 0)$$

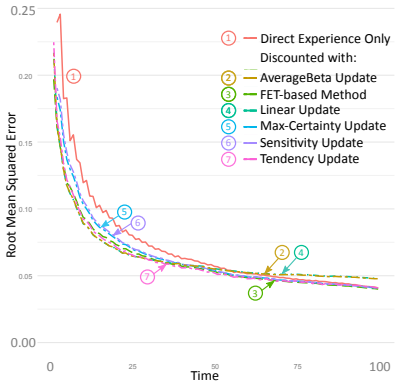
While providing for a smoothing effect, the results did not differ qualitatively from those obtained for fixed $p = 0.5$. This holds for both scenarios; that is, for a recommender misreporting from time step 0 (Figure 12, p. 134) and a recommender behaving honestly first and then misreporting from time step 51 (Figure 13, p. 135)



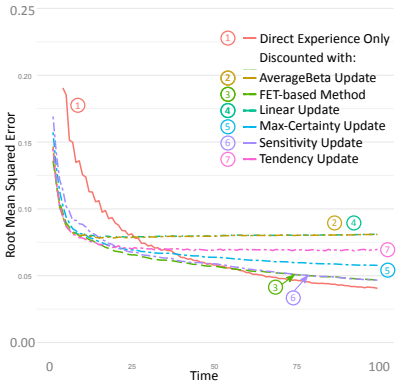
(a) $p = 0.5, \lambda = 1, Offset = 0.05$



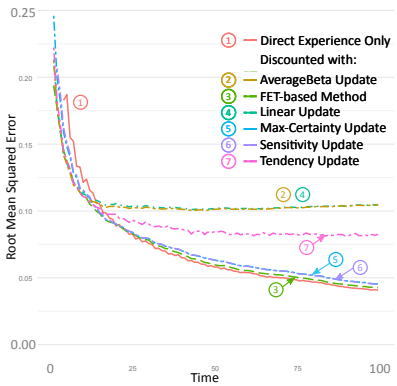
(b) $p = 0.5, \lambda = 5, Offset = 0.05$



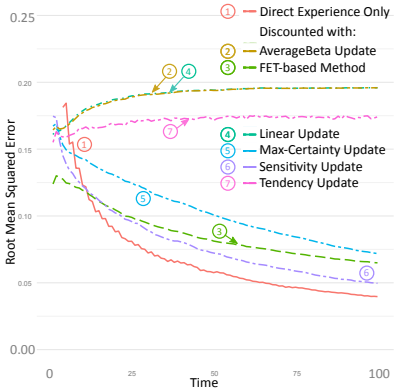
(c) $p = 0.5, \lambda = 1, Offset = 0.10$



(d) $p = 0.5, \lambda = 5, Offset = 0.10$

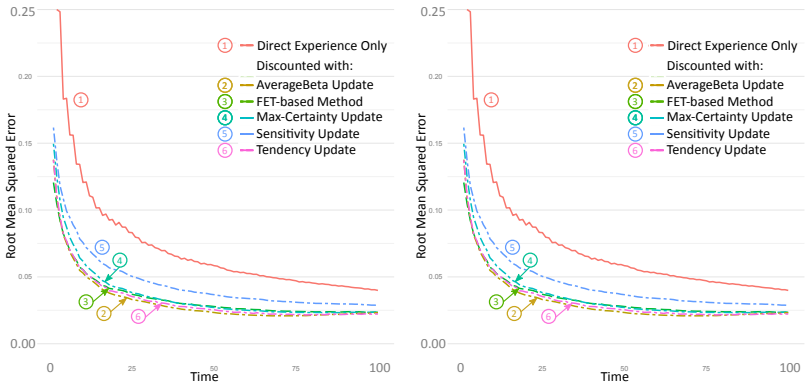


(e) $p = 0.5, \lambda = 1, Offset = 0.25$

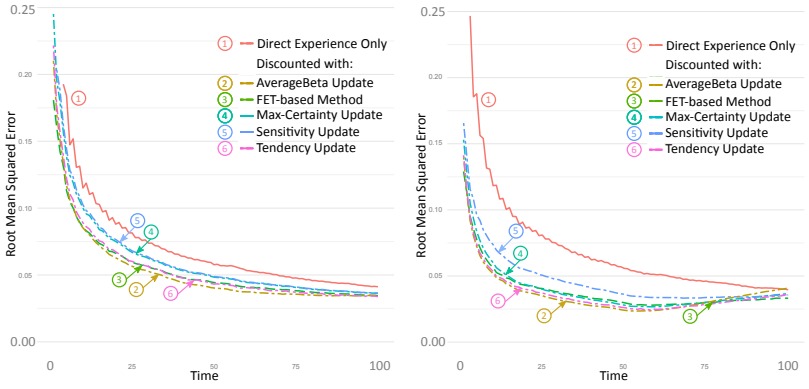


(f) $p = 0.5, \lambda = 5, Offset = 0.25$

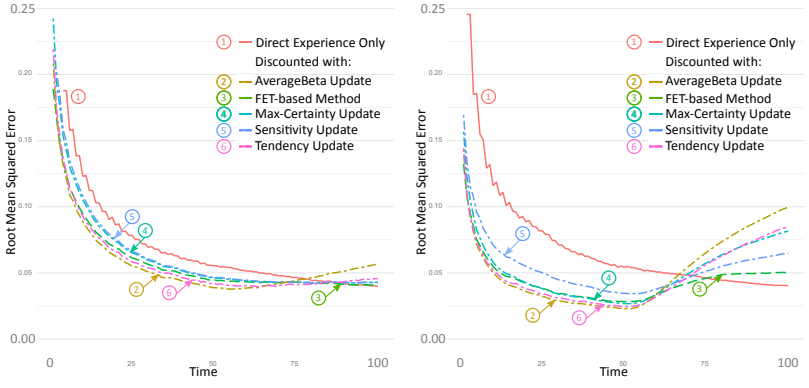
Figure 10: Root Mean Squared Error of the prediction quality under various recommender trustworthiness estimation mechanisms, misreporting recommender, various offsets.



(a) $p = 0.5, \lambda = 1, Offset = 0.05$ (b) $p = 0.5, \lambda = 5, Offset = 0.05$

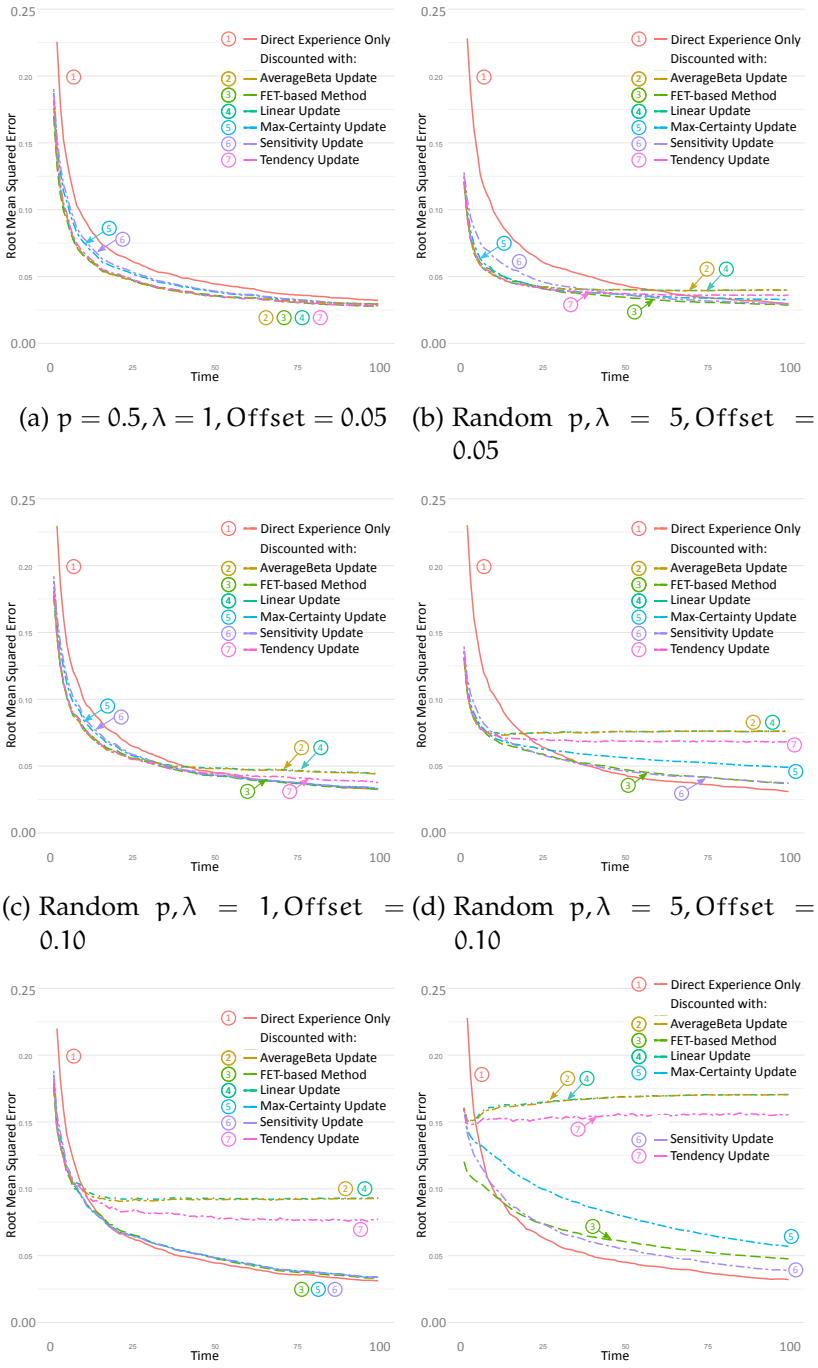


(c) $p = 0.5, \lambda = 1, Offset = 0.10$ (d) $p = 0.5, \lambda = 5, Offset = 0.10$



(e) $p = 0.5, \lambda = 1, Offset = 0.25$ (f) $p = 0.5, \lambda = 5, Offset = 0.25$

Figure 11: Root Mean Squared Error of the prediction quality under various recommender trustworthiness estimation mechanisms, misreporting recommender, various offsets from time step 50.

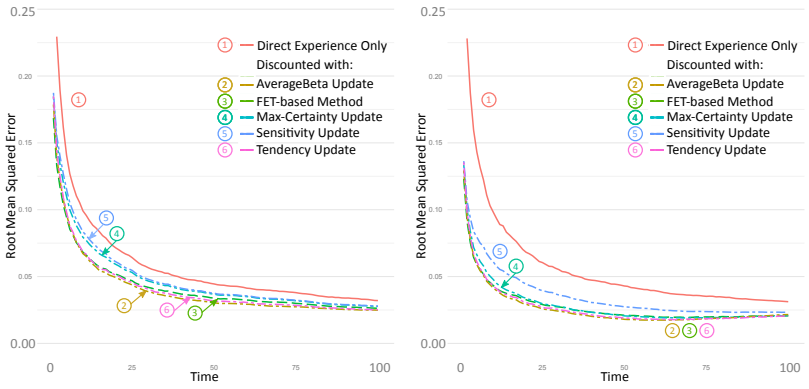


(a) $p = 0.5, \lambda = 1, Offset = 0.05$ (b) Random $p, \lambda = 5, Offset = 0.05$

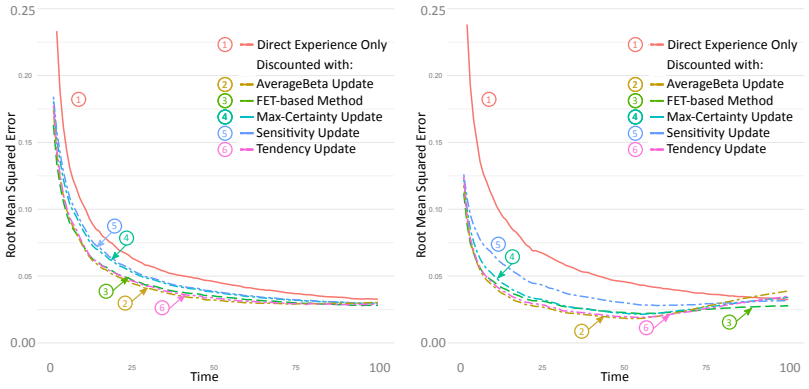
(c) Random $p, \lambda = 1, Offset = 0.10$ (d) Random $p, \lambda = 5, Offset = 0.10$

(e) Random $p, \lambda = 1, Offset = 0.25$ (f) Random $p, \lambda = 5, Offset = 0.25$

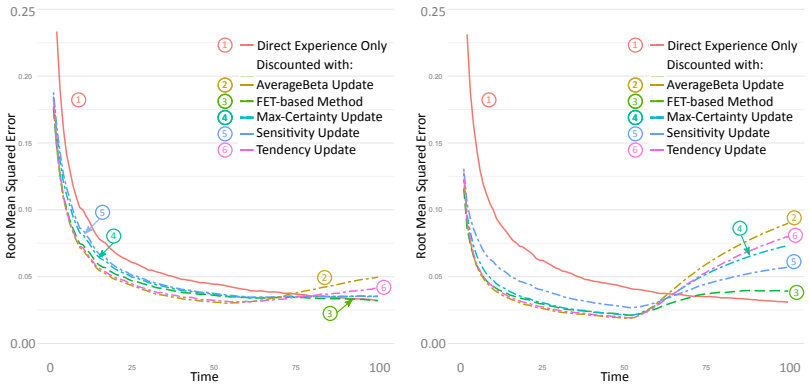
Figure 12: Root Mean Squared Error of the prediction quality under various recommender trustworthiness estimation mechanisms, misreporting recommender, various offsets.



(a) Random $p, \lambda = 1, Offset = 0.05$ (b) Random $p, \lambda = 5, Offset = 0.05$



(c) Random $p, \lambda = 1, Offset = 0.10$ (d) Random $p, \lambda = 5, Offset = 0.10$



(e) Random $p, \lambda = 1, Offset = 0.25$ (f) Random $p, \lambda = 5, Offset = 0.25$

Figure 13: Root Mean Squared Error of the prediction quality under various recommender trustworthiness estimation mechanisms, misreporting recommender, various offsets from time step 50.

SUMMARISING THE SIMULATION RESULTS The simulation results show that those related methods that perform best when assuming honesty/accurate reporting on part of recommender R – *linear*, *average-β* and *tenedency* update – are also those that exhibit performance deficiencies under the assumption of a dishonest/misreporting recommender. The *FET-based* recommender trustworthiness estimation method offers a good mix of performance under honest recommendations and robustness under dishonesty. Additionally, it is not reliant on constant updates, but can be computed independent from scratch at each time step. This eliminates theoretical concerns of the independence of observations when assuming that recommender trustworthiness follows a *Beta*-distributed model.

However, computing the *FET* test statistic is computationally more expensive[15] than executing, for instance, a linear or sensitivity update. This disadvantage can be partially overcome either by using a less computationally expensive test statistic, such as a χ^2 approximation, or by not computing the exact statistic during each time step, but only every n-th time step. For the latter solution, the *FET-based* method may be combined with a momentum term that computes the speed of convergence of the $rmse$; based on the momentum, the frequency of recomputing recommender R's trustworthiness can be controlled. The realisation of such a method is relegated to future work.

If the computational constraints do not permit the use of the *FET-based* approach, the *sensitivity* and *max-certainty* update methods also offer reasonably good performance , while the *tendency* update methods performs surprisingly well given its very simple nature.

4.3.7 *Section Summary*

The previous section extends the consensus and discounting operations of the *CertainTrust* model to the multinomial space, so as to be compatible with the extended *Multinomial CertainTrust* model and afford *CertainTrust* the same *basic* multinomial capabilities as *Subjective Logic*[16] [104, 105]. Furthermore, the section introduces extensions to the fusion operation for aggregating potentially dependent opinions:

- the basic average fusion operation from *CertainLogic* [175] was extended to the multinomial space,
- the weighted average fusion operation from [77] was introduced and extended to the multinomial space, and
- the conflict-aware average fusion operation from [77] was introduced and extended with a statistically accurate, *FET*-based

15 For the computational expense of computing *Fisher's Exact Test*, see [2, 149].

16 *Subjective Logic* incorporates a considerably larger number of specialised operators; these are easily accessible by leveraging the isomorphism from *CertainTrust* to the evidence representation using sufficient statistics of the multinomial samples.

degree of conflict measure, that also allows easy application to multinomial opinions.

By introducing weighting and conflict-awareness, the averaging process of opinions can take into account the fact that the same aggregate can be produced from either very similar or hugely differing opinions. Under basic average fusion, this difference was not considered; under conflict-aware fusion, however, differing opinions have a negative impact on the certainty reported in the resulting fused opinion.

Finally, recommendations are combined with direct experience, using the consensus and discounting operations. Specifically, recommender trustworthiness estimates – as introduced in Section 4.2, p. 100 – are used as discounting factors, thereby enabling the direct comparison of different recommender trustworthiness estimation methods in terms of predictive accuracy. The novel *FET*-based method is shown to perform consistently among the best methods in terms of predictive performance, both under assumptions of honest recommender behaviour and misreporting behaviour.

4.4 LOCAL STATIONARITY AND CHANGE POINT DETECTION

In Chapter 3, one of the major assumptions made when estimating both the trust score ($t_i = \hat{p}_i = \frac{\alpha_i}{\sum_{j=1}^{m} \alpha_j}$) and the certainty values (Definitions 10, 12, 19, and 20), was *stationarity* of the random processes thought to be generating the binomial and multinomial samples. By demanding independent and identically distributed (*iid*) random variables, the stationarity of the random processes allowed for a simple justification of the methods of statistical inference used. This extends in particular to asymptotical arguments, such as the consistency of the estimators. For instance, for *iid* observations, the estimator $\hat{p} = \frac{x}{n}$ of an n-times repeated Bernoulli trial with probability p, that is, for a sample $X \sim \mathrm{Bin}(n, p)$, it holds that $\lim_{n\to\infty} \hat{p} = p$.

Loosely defined, a stationary process is a random process whose statistical properties do not change over time [155]. Applied to the Bayesian trustworthiness estimation task at hand, and the estimators presented in Chapter 3, this has the following implications: In the binomial case, $X \sim \mathrm{Bin}(n, p)$, the estimand parameter p remains unchanged over time. In the multinomial case, $X \sim \mathrm{Mult}(n, \mathbf{p})$, the estimand parameter vector $\mathbf{p} = (p_1, p_2, \ldots, p_m), m \geqslant 2$, remains unmodified. In other words, the stationary model assumes that trustee behaviour does not fundamentally change over time.

However, in the real world, stationarity assumptions frequently do *not* hold.[17,18] Behaviour – for example, of actors in a market environ-

17 Compare, for instance, Granger's seminal work on econometric time series [74].

18 Quoting [155, 191]: *'[...] stationarity and Gaussianity are fairy tales invented for the amusement of undergraduates.'*

ment – tends to change as time passes. In other words, the estimand experiences *concept drift*.

4.4.1 *Concept Drift*

The term *concept* in concept drift refers to the estimand parameter – in the context of trustworthiness estimation, it thus describes a change in the parameter p of the assumed binomial distribution $\mathtt{Bin}(n, p)$ or the parameters $\mathbf{p} = (p_1, \ldots, p_m)$ of the assumed multinomial distribution $\mathtt{Mult}(n, \mathbf{p})$ over time. In a sense, the parameter[19] p is changing, or *drifting*, leading to a non-stationary environment.

The notion of concept drift is primarily related to online *supervised* learning scenarios, where the relation between data input and target variable change over time [63]. Following Bayesian decision theory [48], the posterior probability of an outcome to belong to category $i \in \{1, \ldots, m\}, m \geqslant 2$, given some input data X, that is, $p(\theta_i|X)$, can be described as:

$$p(\theta_i|X) = \frac{p(X|\theta_i) \cdot p(\theta_i)}{p(X)} \tag{15}$$

where $p(\theta_i)$ is the prior probability of category i, $p(X|\theta_i)$ is the category conditional probability function for classes $\{1, \ldots, m\}$, with $p(X) = \sum_{i=1}^{m} p(\theta_i) \cdot (X|\theta_i)$.

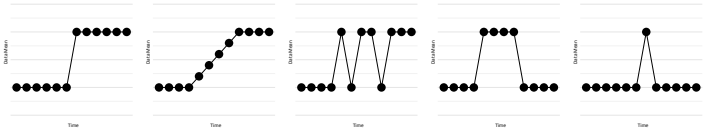
The predictive model presented in Chapter 3 assumes a non-associative learning process, in the sense that no covariate attributes (aside from the history of past experiences) are used in the prediction process. Here, time series information is available for classification. This time series is generated in a reinforcement learning-style feedback loop, modelled by a binomial or multinomial probability distribution. In the following, therefore, a loose interpretation of concept drift will be adopted from [123]. This states simply that under concept drift '*[...] behaviors and tasks change with time*'.

Recall from Chapter 3, that $\mathtt{Posterior} \propto \mathtt{Likelihood} \times \mathtt{Prior}$. In fact, from Bayes' Theorem given in Equation 15, the posterior is *equal* to the likelihood function, $\theta \mapsto p(X|\theta_i)$, multiplied by the prior probability $p(\theta_i)$ if it is normalised by the probability of the data $p(X)$. The likelihood function is determined by the data generating process. The prior is determined by the convenient choice of an appropriate *conjugate prior*[20]. For the intents and purposes of the trustworthiness assessment task, likelihood function and prior can thus be considered as fixed.

Consequently, the only constituent of Equation 15 that may exhibit variation over time is the probability distribution of the data, $p(X)$.

19 In the following, unless specifically remarked upon, the single parameter p of the binomial model and the parameter vector $\mathbf{p}$, will not be differentiated.

20 On the concept of conjugate priors, see [167].



(a) sudden (b) incremental (c) gradual (d) reoccuring (e) outlier

Figure 14: Patterns of changes in data over time (outlier not concept drift) [63] (for larger figures, see Appendix E, p. 241).

As discussed in Chapter 3, the distributional family of the sample X is known, belonging either to the binomial or multinomial family of distributions. The objective was the estimation of the shape parameter(s) of the specific, stationary distribution generating the data sample $X = \{x_1, x_2, \ldots, x_n\}$.

Under assumptions of non-stationarity, the shape parameter p of the generating binomial distribution may change over time, from the initial value of the shape parameter, p, to a new value p'. This change in $p(X)$ will manifest itself in the sample X in proportion to the magnitude of the change in p. In particular, the change of p over time is evidenced by a significant change in the data mean, that is $\hat{p} = \frac{\sum_{i=1}^{n} x_i}{n}$. [63] categorise four different types of drift in p:

- *sudden/abrupt* (Figure 14a): an abrupt switch from p to p' at a fixed point in time, t, so that $\{x_1, \ldots x_{t-1}\} \sim \mathrm{Bin}(t, p)$ and $\{x_t, \ldots, x_n\} \sim \mathrm{Bin}(n-t, p')$.

- *incremental* (Figure 14b): a change occurring over many intermediate steps, so that for $0 < t' \leqslant t \leqslant t'' < n$ it holds: $\{x_1, \ldots x_{t'-1}\} \sim \mathrm{Bin}(t', p)$, $\{x_{t''}, \ldots, x_n\} \sim \mathrm{Bin}(n-t'', p')$ and an incremental, monotonous function of p for $t' < t < t''$, so that $f(p) \in [p, p']$ if $p' > p$ and $f(p) \in [p', p]$ if $p > p'$.

- *gradual* (Figure 14c): a change that sees an overlap of concepts, so that for an interval of time $[t, t']$, data may be generated by either the original distribution with shape parameter p or the new distribution with shape parameter p'.

- *reoccuring* (Figure 14d): a switch where the original distribution with shape parameter p reoccurs after an interval of time $[t, t']$, so that $\{x_1, \ldots x_{t-1}\} \sim \mathrm{Bin}(t, p)$, $\{x_t, \ldots x_{t'-1}\} \sim \mathrm{Bin}(t'-t, p')$ and $\{x_{t'}, \ldots x_n\} \sim \mathrm{Bin}(n-t', p)$.

As with other online learning scenarios, the different forms of concept drift listed above can occur in trust assessment. Therefore, trust models have to make predictions from evolving data with unknown dynamics. Leveraging the assumption that the most recent data is the most informative for the prediction process [63, 46], state-of-the-art trust models employ gradual forgetting and windowing mechanisms.

4.4.2 *Gradual Forgetting*

Gradual forgetting, or *ageing* of evidence, is the basic concept for handling concept drift in computational trust models.; as such, models incorporating gradual forgetting include, for instance, [29, 107, 108, 173, 202]. The implementation of gradual forgetting is achieved by weighting the individual experiences contained in sample $\tilde{X} = (\vec{x}_1, \vec{x}_2, \vec{x}_3, \ldots, \vec{x}_{n-1}, \vec{x}_n)$ in $0-1$ random vector form based on their age. This follows from the intuition that the importance of an experience in the sample should decrease with age [63].

Definition 35 (Gradual Forgetting (Ageing))**.** Let $a \in [0;1]$ be a *fading factor* and $\tilde{X} = (\vec{x}_1, \vec{x}_2, \vec{x}_3, \ldots, \vec{x}_{n-1}, \vec{x}_n)$ be a sample in $0-1$ *random vector* form with dimension $m \times n$. Let $\tilde{X}$ be ordered according to the age of its components, so that $\vec{x}_1$ is the oldest and $\vec{x}_n$ the most recent component, i.e., $\tilde{X}$ represents a time-series. The *aged* sample $\tilde{X}_{n,a}$ is computed as:

$$\tilde{X}_{n,a} = (a^{n-1} \cdot \vec{x}_1, a^{n-2} \cdot \vec{x}_2, a^{n-3} \cdot \vec{x}_3, \ldots, a^1 \cdot \vec{x}_{n-1}, \vec{x}_n)$$

For the $m \in \mathbb{N}, m \geqslant 2$ row sums in $\tilde{X}_{n,a}$ it follows that for $1 \leqslant i \leqslant m$:

$$\tilde{\alpha}_i = \sum_{j=1}^{n} a^{n-j} \cdot x_{i,j}$$

The $\tilde{\alpha}_i$'s from Definition 35 now provide the parameters for the trustworthiness estimation with an *aged* sample that gives less importance to older data. Thus, the trust scores $t_1, t_2, \ldots, t_m$ that are computed with aged data as expectation values of Beta or Dirichlet posterior distributions take the following form:

$$t_i = \frac{\tilde{\alpha}_i}{\sum_{j=1}^{m} \tilde{\alpha}_j}, \text{for } i \in \{1, 2, \ldots, n\} \tag{16}$$

Correspondingly, because the amount of information thought to be contained in the sample is reduced by the ageing assumption, certainty estimates must also be adjusted. The certainty estimates should therefore be computed by substituting $\sum_{i=1}^{m} \tilde{\alpha}_i$ for n in the various certainty estimators (Definitions 10, 12, 19 and 20). Thus, the general instantiation for a certainty estimator using an aged sample is:

$$C(\tilde{\alpha}_1, \tilde{\alpha}_2, \ldots, \tilde{\alpha}_m; \sum_{i=1}^{m} \tilde{\alpha}_i)$$

AGEING STORED SUFFICIENT STATISTICS Recall that the aggregates over the $m \in \mathbb{N}$ different categories in the sample, that is, $\alpha_1 = \sum_{j=1}^{n} x_{1,j}, \ldots, \alpha_m = \sum_{j=1}^{n} x_{m,j}$, are sufficient statistics of the sample. For $1 < t < n$, let $\alpha_i(t) = \sum_{j=1}^{t} x_{i,j}$ be the sum over the first $t \in \mathbb{N}$ entries in the i-th category. When using ageing with fading

factor $a \in [0;1]$ and storing only aggregate data at each time interval t, i.e., $\alpha_i(t)$, and maintaining only the most current $\vec{x}_{(t+1)}$, the gradual forgetting mechanism each time, the sufficient statistics can be updated in a convenient manner:

$$\alpha_j(t+1) = x_{j,t+1} + a \cdot \alpha_j(t)$$

PERIODS OF INACTION Ageing raises an issue related to periods of inaction, when no new evidence is collected but time passes nonetheless. Assuming discrete time intervals in which an interaction leading to evidence can occur, a period of inaction can easily be accommodated by relaxing the requirements of the $0-1$ *random vector* representation of the sample. In its standard form (see Tables 3, p. 78 and 4, p. 80), a sample in this representation consists of m-dimensional column vectors containing *exactly one* 1-element and $m-1$ 0-elements. For a sample of length $n \in \mathbb{N}$, it thus follows that $\sum_{i=1}^{n}\sum_{j=1}^{m} x_{i,j} = n$. In order to accommodate periods of inaction, let us formally define a period of inaction as a column vector in sample X containing *only* 0-elements. In this representation, it follows that a sample X of length $n \in \mathbb{N}$ encodes $\sum_{i=1}^{n}\sum_{j=1}^{m} x_{i,j} \leqslant n$ periods during which an interaction occured and $n - \sum_{i=1}^{n}\sum_{j=1}^{m} x_{i,j} \geqslant 0$ periods of inaction. Table 7, p. 142 shows a sample containing a period of inactivity encoded as 0-element vector $\vec{x}_3$. Also note, that the trust estimate computation had to be adjusted to $t_u = \hat{p}_u = \frac{1}{\sum_{i=1}^{n}\sum_{j=1}^{m} x_{i,j}} \cdot \sum_{i=1}^{n} x_{u,i}$ in order to account for the (possible) occurrence of 0-element vectors. In order to avoid a division by zero, at least a single non-zero vector should be present in the sample. Otherwise, the default instantiation of *CertainTrust* is used, i.e., $t = 0.5$, $c = 0$.

This representation serves as an auxiliary representation in order to account for inaction. It simply enables an easy application of ageing per Definition 35, which assumes discrete time steps.

LIMIT ON EVIDENCE THROUGH AGEING Ries [173] noted a narrowing of the interval that can contain the expectation values of the Beta distributed posteriors. This narrowing is owed to the increased impact of the $\mathrm{Beta}(1,1)$ prior in relation to the data as the data decays over time. The amount of evidence that can be obtained when ageing with fading factor $a \in [0;1]$ can be expressed as $\sum_{i=1}^{m} \tilde{\alpha}_i$ (see Definition 35, p. 140). This amount of evidence is limited and can be expressed as a geometric sum[21]:

$$\sum_{i=1}^{m} \tilde{\alpha}_i = \sum_{j=1}^{n} (a^{n-j} \cdot \|\vec{x}_j\|) \leqslant \sum_{j=0}^{n-1} a^j = \frac{1-a^n}{1-a} \tag{17}$$

21 The inequality stems from possible 0-element vectors in the sample. If no such vectors are in the sample, the $\leqslant$-relation becomes an equality.

Category	Sample $\tilde{X}$ (w/ Length n)						Trust Estimate
	$\vec{x}_1$	$\vec{x}_2$	$\vec{x}_3$	$\dots$	$\vec{x}_{n-1}$	$\vec{x}_n$	t
Cat_1	1	0	0	$\dots$	1	0	$\hat{p}_1 = \frac{1}{\sum_{i=1}^{n}\sum_{j=1}^{m} x_{i,j}} \cdot \sum_{i=1}^{n} x_{1,i}$
Cat_2	0	1	0	$\dots$	0	0	$\hat{p}_2 = \frac{1}{\sum_{i=1}^{n}\sum_{j=1}^{m} x_{i,j}} \cdot \sum_{i=1}^{n} x_{2,i}$
$\vdots$				$\ddots$			$\vdots$
Cat_m	0	0	0	$\dots$	0	1	$\hat{p}_m = \frac{1}{\sum_{i=1}^{n}\sum_{j=1}^{m} x_{i,j}} \cdot \sum_{i=1}^{n} x_{m,i}$
$\sum_{\{Cat_1,\dots,Cat_m\}} \vec{x}_i$	1	1	0	$\dots$	1	1	$\sum_{j=1}^{m} \hat{p}_j = 1$

Table 7: Multinomial Trust Assessment in 0-1 Random Vector Form Encoding Periods of Inaction.

Thus, even assuming an infinite sample length $n \to \infty$, the total amount of evidence is still finite, as

$$\lim_{n\to\infty} \frac{1-a^n}{1-a} = \frac{1}{1-a} \tag{18}$$

This issue is related to the fact that with limited amounts of information, only a limited degree of certainty in the estimate can be achieved. Thus, the prior, in the case of [173] and most other trust models a Beta(1,1) prior, is not entirely marginalised by the data. However, this is not just an issue with trust models, but present in Bayesian estimation in general, in particular when ageing is applied. As a practical solution, [173] proposes to dynamically adjust the prior and limit the amount of information required to reach a certainty value of $c = 1$ by fading out the prior. The method for doing so has already been extended from the basic *CertainLogic* in Section 3.1.7 for the binomial model with arbitrary certainty functions and in Section 3.2.7 for the novel *Multinomial CertainTrust* extension.

EXTENDED IMPLICIT AGEING The implicit ageing process by normalisation proposed in [173] for *CertainTrust* can be generalised to the multinomial case. For this, assume a fixed $N \in \mathbb{N}$ to represent the minimum number of experiences required to reach a certainty value of $c = 1$. The value of N can, for instance, be computed according to the mechanism proposed in Section 3.2.7, p. 89. Let $X = (\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_n)$ be a sample in $0-1$ *random vector* representation with dimension $m \times n$. Then the normalisation analogue to [173] is given as:

$$\mathrm{norm}(X) = \begin{cases} (\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_n) & \text{if } \sum_{i=1}^{n} \|\vec{x}_i\| \leqslant N \\ \left(\frac{N\cdot\vec{x}_1}{\sum_{i=1}^{n}\|\vec{x}_i\|}, \frac{N\cdot\vec{x}_2}{\sum_{i=1}^{n}\|\vec{x}_i\|}, \ldots, \frac{N\cdot\vec{x}_n}{\sum_{i=1}^{n}\|\vec{x}_i\|}\right) & \text{else.} \end{cases}$$

where $\|\vec{x}_i\|$ is the Euclidian vector norm. For the given sample in $0-1$ *random vector* representation, it holds that $\sum_{i=1}^{n} \|\vec{x}_i\| = \sum_{i=1}^{n} \sum_{j=1}^{m} x_{i,j}$, that is, the sum of elements in X.

The presented mechanism of gradual forgetting and the consequent refinements, such as implicit ageing, however, do not address two pertinent questions: Is the application of ageing warranted in a given application and what should the fading factor $a \in [0;1]$ be instantiated as? Both questions relate to the fact that ageing limits the amount of evidence that can be collected, which in turn impacts the prediction results. Therefore, ageing should be complemented by a technique that indicates whether or not the underlying distribution within the sample X has changed over time, and if so, at which points. That is, over which intervals in the sample the distribution can be assumed to be stationary. This can be achieved by applying methods for *change point detection* to trustworthiness assessment. The change point detection model applied to trustworthiness estimation, which adopts the

methodology from Ross et al. [178], is primarily geared towards a binomial model. However, it is easily extended to the multinomial case (see, Section4.4.5, p. 150).

4.4.3 *Change Point Detection*

In general, change point detection encompasses techniques and methods for the identification of points in time (or small intervals in time) at which the estimand parameter of the distribution of a time series changes. By identifying change points in the distribution of the data, the dynamics of the data generating process are explicated [63]. This, in turn, is useful for both correctly determining the trust estimate $t = \hat{p}$ and its concordant certainty estimate c.

Methods of change detection are mature statistical procedures, of particular interest in manufacturing and quality control. Statistical process control[22] has been used for a number of decades, going back to Shewhart's work [184] on so-called control charts while at Bell Laboratories in the 1920s [15]. In fact, Shewhart's goal was similar to the challenge posed in trustworthiness estimation under non-stationarity[23]: Shewhart's original control chart, now called a p-chart, was designed for monitoring the fraction of defective items in a manufacturing process [22]. Other control chart approaches have superseded the original p-chart in terms of effectiveness to detect small to moderate shifts in the fraction of defectives, such as the Binomial and Bernoulli Cumulative Sum (CUSUM) control charts [22, 170] or the FETCPM method [178]. Within the context of reputation systems, Yang et al. [206] have applied change detection methods to detect collaborative reputation attacks in a centralised manner, albeit with assumptions of Gaussianity and not for coping with non-stationary recommender behaviour.

BINOMIAL SAMPLES WITH CHANGE POINTS The structure of the $m \times n$-dimensional data sample $X = (\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_n)$ in $0-1$ random vector form can be considered as m different row samples. Each of these m samples $Y_i = (x_{i,1}, x_{i,2}, \ldots, x_{i,n}), 1 \leqslant i \leqslant m$, is a discrete-time sequence of realisations of independent Bernoulli distributed random variables. Under the assumption of local stationarity, the parameter p of the data generating Bernoulli process may change at any of the $n \in \mathbb{N}$ data points in the sample Y_i. Therefore, assuming that the Bernoulli parameter p_t is constant but unknown between change

22 For an overview, see [159].

23 Recall that an assumption of local-stationarity is still made.

points, the distribution of $X_{i,j}$, which denotes the random variable realised by observation $x_{i,j}$, can be written as

$$X_{i,j} \sim \text{Bernoulli}(p_{i,j}), p_{i,j} = \begin{cases} p_{i,0} & \text{if } j \leqslant \tau_1 \\ p_{i,1} & \text{if } \tau_1 < j \leqslant \tau_2 \\ p_{i,2} & \text{if } \tau_2 < j \leqslant \tau_3 \\ \dots \end{cases}$$

where each value of τ_{idx} represents a change point [178]. A change point detector is tasked with identifying changes in the different $p_{i,j}$ as quickly as possible, that is, its task is determining the change points.

The nature of data generation in Bayesian trust estimation, however, places additional requirements on the change point detection method. In particular, samples do not have a fixed size, but rather may arrive as data streams without a fixed length. Additionally, the computational and memory demands of the change point detector should be moderate in order to remain feasible. This, in conjunction with the discrete nature of the data, limits the choice of change point detectors to methods that are capable of sequential analysis of attribute data under 100 per cent inspection.

SEQUENTIAL CHANGE POINT DETECTION In sequential change point detection, the sample is analysed in sequence as a new data point arrives. The change detector is sequentially applied to the sample until a change in the probability of the data generating (Bernoulli) process is detected. After such a change point has been identified, the sequential change point detector is reset, so as to detect the next change point. In this setting, the problem of change point detection is solved in a successive manner, in which a change from a pre-change probability value θ_0 to a post-change probability θ_1 is signalled when a change point is identified. After the identification of the change point and the subsequent reset of the detector, what was formerly the post-change probability becomes the new pre-change probability and detection continues.

Sequential change point detectors rely on hypothesis testing. For a data point k in a sequence, a simple two-sample hypothesis test can be used to check whether a change point occurs at $\tau = k$, with the null-hypothesis, h_0, that there is no change at k [63, 178]. The underlying assumption is that for a change at k and a time-series of data points indexed $1 < k < n$, the sequence with indices 1 to $k-1$ and the sequence with indices k to n have a significantly different probability of exhibiting certain subsequences. Typically, the hypothesis is

tested using the Sequential Probability Ratio Test *(SPRT)* [193], with the test statistic taking the form of:

$$T_k^n = \log \frac{P(x_k, \ldots, x_n|\theta_1)}{P(x_k, \ldots, x_n|\theta_0)} = \sum_{i=k}^{n} \log \frac{\theta_1[x_i]}{\theta_0[x_i]} = T_k^{n-1} + \log \frac{\theta_1[x_n]}{\theta_0[x_n]}$$

with a change being signalled if T_w^n exceeds a predefined threshold [63]. Note, that the pre-change probability, θ_0, is assumed to be known.

The Bernoulli CUSUM change point detector with 100 per cent inspection [22, 170] is one of the state-of-the-art methods for change point detection under the given conditions. However, in order to operate efficiently, this detector requires the pre-change probability, θ_0, to be exactly known, due to its reliance on the SPRT. Under misspecification of θ_0, performance of this detector is significantly impaired [23, 178]. In trust assessment, as a feedback-based learning task, however, the pre-change probability is generally not known a-priori, thereby potentially limiting the usefulness of the Bernoulli CUSUM detector.

A more recent sequential change point detector has been introduced in [178]. Instead of relying on the SPRT, it uses Fisher's Exact Test (Definition 26, p. 108) *(FET)* [2, 178] in the context of the change point model *(CPM)* framework introduced in [87][24]. The FET does not rely on Wald's Gaussian approximations used in the SPRT, permitting its application in scenarios with small sample sizes. Neither does it depend on the true, yet unknown, pre-change probability θ_0. Rather, since we are only dealing with discrete, binary data points, the FET statistic is constructed from a combinatorics argument, that reasons over the distribution of *failures* (i.e., $x_i = 0$) in the sample.

4.4.4 *FETCPM Change Point Detector*

In the following, the FETCPM detector is described according to Ross et al. [178], from which the fundamental *FET*-based change point detection model is largely reproduced. The detector relies on the FET [2, 56], which has already been used in previous sections and defined in Definition 26, p. 108. To recapitulate briefly, the FET tests whether two binomial samples are generated by the same Bernoulli process. In its two-sided variant, the FET gives the probability that a null-hypothesis, h_0 (*the samples originate from the same Bernoulli process*), is true. Within the context of change point detection, h_0 can be reformulated as: *there are no change points within a given sample*. The following Defintion 36 puts the FET into the context of Change Point Detection problems:

24 Hence, this particular change point detector will, in the following, be referred to as *FETCPM*.

Definition 36 (Fisher's Exact Test [56] in the context of Change Point Detection [178])**.** Let $X = (x_1, x_2, \ldots, x_n)$ be a sequence of realisations of Bernoulli distributed random variables X_i. Let $k, 1 < k < n$ be a point in the sequence X that splits the sequence into two samples $x_1, \ldots, x_k$ and $x_{k+1}, \ldots, x_n$. Furthermore, let the null-hypothesis, h_0, be that there are no change points in X. Under h_0, all X_i are identically distributed with $P(X_i = 1) = \theta_0$ and $P(X_i = 0) = 1 - \theta_0$. Let S_n be a random variable defined as the sum of failures in a sequence of length n:

$$S_n = \sum_{i=1}^{n} (X_i = 0)$$

Correspondingly, let S_k be a random variable of the sum of failures in the first sample $x_1, \ldots, x_k$. Then, conditional on S_n being realised as $S_n = s_n$, the probability that $S_k = s_k$ follows a hypergeometric distribution

$$P(S_k = s_k | S_n = s_n) = \frac{\binom{s_n}{s_k}\binom{n-s_n}{k-s_k}}{\binom{n}{k}} \tag{19}$$

This, in turn leads to the one-sided p value of the FET, that is, the probability that there are s_k or less failures in the first k observations under h_0 [178]:

$$p_{k,n} = \sum_{i=1}^{k} P(S_k = s_i), p_{k,n} \in [0; 1]$$

The one-sided version of the FET in Definition 36 detects an increase in the estimand parameter, that is, increases in the parameter p of the underlying Bernoulli distribution. In order to detect decreases, the sufficient statistic S_n (and correspondingly S_k) has to be replaced by the sum of successes:

$$R_n = \sum_{i=1}^{n} (X_i = 1)$$

In the following, the formalisation will assume the detection of increases in the parameter and hence use the S_n statistic. This is done for reasons of convenience, as the reformulation for the statistic R_n is trivial.

Change detection based on the FET entails that, should a change be detected, the null-hypothesis is rejected. Let $F_{k,n} = 1 - p_{k,n}, F_{k,n} \in [0; 1]$; then the statistic $F_{k,n}$ can simply be tested against an appropriately chosen threshold $h_{k,n}$ and if $F_{k,n} > h_{k,n}$, a change can be signalled [178]. Since it is not known a-priori which point k in the sequence is the change point, a change point τ can be estimated as $\hat{\tau}$

by computing $F_{k,n}$ for every $1 < k < n$ and reporting the maximum value [178], that is:

$$F_n = \max_{1<k<n} F_{k,n}$$

If $F_n > h_n$ for some appropriate h_n, then h_0 can be rejected and the point k at which $F_{k,n}$ is maximised declared to be a suitable estimator $\hat{\tau}$ of the change point τ. The algorithm is given is pseudocode in Algorithm 3.

In the case of trustworthiness estimation, the observations, in the form of experiences, arrive in discrete time as part in what can be considered a stream. That is, the number of observations is not fixed and change detection is done in a sequential manner, necessitating a recompilation of the test statistic F_n for each new observation. However, this can be done in a computationally efficient manner [178], as long as the threshold values h_n are chosen appropriately.

DETERMINING $\{h_n\}$ Each sequential computation of F_n requires a corresponding choice of threshold parameter h_n. The threshold parameter for sequential change point detectors, such as FETCPM or CUSUM, is usually determined on the basis of the zero-state Average Run Length (ARL_0), which is the expected time between false alarms. In this context, a false alarm is constituted by the detector signalling a change when in fact no change has occurred. For this, the probability of a false alarm should be bounded, so that for a user-specified probability value α [178]:

$$P(F_n > h_n | F_{n-1} \leqslant h_{n-1}, \ldots, F_1 \leqslant h_1, p_n = \theta_0) = \alpha$$

The ARL_0 is then defined as $\frac{1}{\alpha}$. Form this, a sequence $\{h_n\}$ of threshold values for sequential testing can be pre-computed and stored as a look-up table. [178] provide such a table for the FETCPM that bounds the ARL_0 conservatively, so that $ARL_0 \leqslant \frac{1}{\alpha}$.

SMOOTHING Since the criterion of $ARL_0 \leqslant \frac{1}{\alpha}$ is already conservative, [178] recommend smoothing the test statistic F_n in order to reduce the overall conservativeness of the FETCPM. The new smoothed statistic is defined by [178] as:

$$\begin{aligned} Y_{1,n} &= F_{1,n} \\ Y_{k,n} &= (1-\lambda) \cdot Y_{k-1,n} + \lambda \cdot F_{k,n} \\ Y_n &= \max_{1<k<n} Y_{k,n}, \lambda \in [0;1] \end{aligned}$$

A value of $\lambda = 0.3$ appears an appropriate choice, according to the performance evaluations of different λ-values in [178].

RECURSIVE COMPUTATION AND CUMULATION Computing F_n for each new observation would normally entail re-computing *each*

Data: Binomial time series $X = (x_1, x_2, \ldots, x_n)$
Pre-computed threshold value h_n
Result: Change point estimate $\hat{\tau}$

// Initialize parameters
n=length(X);
$\hat{\tau} = 0$;
$F_n = 0$;
$s_n = sum\ of\ failures$ in X;
$s_k = 0$;
$k_0 = 0$;
// Compute the test statistic
for k *in* $1 : n$ **do**
  $s_k = sum\ of\ failures$ in $(x_1, \ldots, x_k)$;
  $F_{k,n} = 1 - \frac{\binom{s_n}{s_k}\binom{n-s_n}{k-s_k}}{\binom{n}{k}}$;
  if $F_{k,n} > F_n$ **then**
    $F_n = F_{k,n}, k_0 = k$
  end
end
// Compare the test statistic to threshold
if $F_n > h_n$ **then**
  $\hat{\tau} = k_0$
end
return $\hat{\tau}$

Algorithm 3: FETCPM Change Point Detection according to [178]

$F_{k,n}$, involving the expensive computation of binomial coefficients. That is, computing the entire algorithm that is given in Agorithm 3 in its entire whenever only one new observation is added to the time series X.

However, an algebraic manipulation exploiting the high level of correlation between successive $F_{k,n}$ statistics permits a recursive computation of these values [178]. The recursion takes the following form [178]: Let $d_{k,n}$ be the hypergeometric distribution from the FET (Equation 19, p. 147), i.e.:

$$d_{k,n} = P(S_k = s_k | S_n = s_n) = \frac{\binom{s_n}{s_k}\binom{n-s_n}{k-s_k}}{\binom{n}{k}}$$

Thus, the individual $F_{k,n}$ are computed as:

$$F_{k,n} = 1 - \sum_{i=1}^{k} d_{d_{i,n}}$$

Then $d_{k,n}$, and consequently $F_{k,n}$ and $Y_{k,n}$, can be computed in a recursive manner:

$$d_{k,n+1} = \begin{cases} \frac{d_{k,n} \cdot (s_n+1)(n-s_n+1)^2}{(k+1-s_k)(n-s_n-k-s_k+1)(n+1)} & \text{if } X_{n+1} = 1 \\ \frac{d_{k,n} \cdot (n-s_n+1)^2}{(n-s_n-k-s_k+1)(n+1)} & \text{if } X_{n+1} = 0 \end{cases}$$

$$d_{k+1,n} = \begin{cases} \frac{d_{k,n} \cdot (s_n - s_k)(k+1)}{(s_k+1)(n-k)} & \text{if } X_{n+1} = 1 \\ \frac{d_{k,n} \cdot (n - s_n - k + s_k)}{(k - s_k + 1)(n-k)} & \text{if } X_{n+1} = 0 \end{cases}$$

In order to limit the memory and computation demands that arise from sequentially monitoring, which are growing over time, [178] propose a windowing and cumulation scheme that does not affect performance significantly. This approach includes computing the $F_{k,n}$ statistic over a window of the most recent w observations, so that only the points $x_{n-w+1}, \ldots, x_n$ are being tested for a change point, thereby making both computational and memory cost constant for each newly arriving observation. At the same time, the observations outside the window, that is, $x_1, \ldots, x_{n-w}$, are cumulated into a new sufficient statistic $s_{n-w} = \sum_{i=1}^{n-w}(X_i = 0)$ and S_k can be defined accordingly as:

$$S_k = s_{n-w} + \sum_{i=n-w+1}^{k} (X_i = 0), n - w < k < n$$

4.4.5 *Applying the FETCPM Change Detector in Trustworthiness Assessment*

As the FETCPM change point detector can operate on cumulated data, such as the sufficient statistics R_n and S_n (Equation 19, p. 147), as well as directly on the time series representation, it is easily applicable to both the compact standard *CertainTrust* formulation, as well as samples in $0-1$ random vector form as used in the previous sections of this thesis. The FETCPM detector, as described above, monitors for changes in a *binomial* proportion. Thus, in the case of trust assessment in the binomial case, the application of FETCPM is straightforward; one simply monitors the sum of failures for increases in the parameter p of the Bernoulli distributed process generating the observations, and the sum of successes for decreases in p, as well as a window of adequate size w. Once a change is detected by the change detector, the observations made at time points before the signalled change point are simply discarded. This is justified, because those observations are assumed to be generated by a different Bernoulli process than the one currently active. As a consequence, change detection implements *abrupt forgetting* conditional on the thresholded statistic of the change detector. Initialising the FETCPM change detector requires observing a minimum number of data points that are assumed to be *independent and identically distributed (iid)*; the *R*-implementation of FETCPM in [177] recommends 20 observations.

Up to this point, only *binomial* change point detection has been In order to monitor multinomial proportions in trust models such as *Multinomial CertainTrust*, individual change detectors are run on *each* of the $m > 2$ proportions in an m-dimensional trust estimation task

(see, Section 3.2.1, p. 77)[25]. If a change detector detects a change in the marginal distribution for any of the $m > 2$ different categories, a change is signalled and *all* change detectors are reset. The basic assumption made so far for multinomial trust assessment was that the sample follows a discrete multinomial proportion with a very simple correlation structure between the different possible outcomes of an observation. That is, the individual observations are independently distributed and the $m > 2$ different categories are exclusive and exhaustive.

Therefore, discarding the notion that the observations are also identically distributed, but rather assuming local stationary and the possibility of change points instead, consider each category to be generated by a Bernoulli process with *probability of success* p_i for the i-th category (and correspondingly, a *probability of failure* of $1 - p_i$). Because the categories are by design exhaustive and exclusive, it follows that $\sum_{i=1}^{m} p_i = 1$. Therefore, a significant change, according to the FET statistic, in one p_i in the positive direction will manifest itself in the cumulated sum of successes over all other categories. A significant change in the negative direction, that is a decrease in p_i will be observable from the number of success of the i-th category. Hence, for each category two statistics have to be maintained:

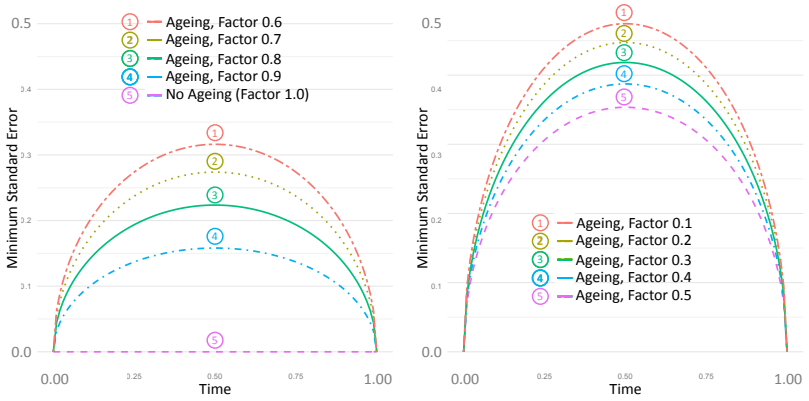
- the sum of successes for the i-th category, $R_n^i = \sum_{j=1}^{n} x_{j,i}$, as well as
- the sum of failures for that category, which corresponds to the sum of success for all but the i-th category, $S_n^i = \sum_{k=1;k\neq i}^{m} \sum_{j=1}^{n} x_{j,k}$.

Depending on the choice of window size, w, the *detection time*, det, at which a change point is signalled and the estimate $\hat{\tau}$ of this change point max not coincide. That is, $det \geqslant \hat{\tau} \geqslant k_d - w$. At time $det + 1$, the reset statistic is therefore computed starting at time $\hat{\tau}$, instead of det if $det \neq \hat{\tau}$. Trust and certainty estimates (Chapter 3) are also restarted from time $\hat{\tau}$.

4.4.6 *Evaluation: Complementing and Comparing FETCPM with Ageing*

Ageing, in the form of gradual forgetting of older information (Section 4.4.2, p. 140) is employed to increase the responsiveness of an estimator to concept drift. This is achieved by essentially bounding the amount of evidence that can be collected (Equations 17 and 18, p. 141), thereby reducing the mass older information has in the cumulation relative to more recent information and consequently increasing the momentum of change in the predictor. However, the equality of the estimate $\hat{p}$ and the estimand p is an asymptotical argument, that

25 This is a straightforward implementation for multinomial and multiattribute change detection problems appropriate under the specific circumstances of trust assessment; for a review of various other methods, see [192]



(a) Ageing factors $a = 0.6$ to $a = 1$ (b) Ageing factors $a = 0.1$ to $a = 0.5$

Figure 15: Lower bound of the Standard Error for varying parameter values p, depending on ageing factors.

is, it only holds for an infinite amount of evidence under stationary conditions: $\lim_{n\to\infty} \hat{p} = p$.

As such, the confidence that an estimate is correct is bounded by the amount of information that can be collected. For an ageing factor of $a \in [0; 1[$, the amount of evidence is bounded by:

$$\frac{1}{1-a}$$

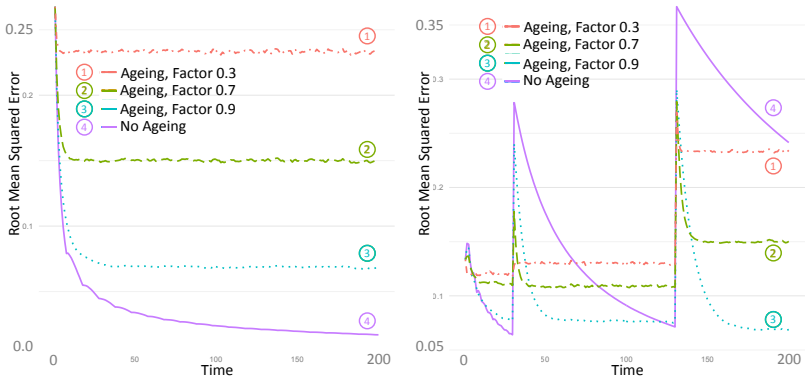
In terms of the Standard Error, this leads to a minimum Standard Error, in a binary sequence, of:

$$\sqrt{p \cdot (1-p) \cdot (1-a)}$$

Ageing trades off the maximum achievable accuracy of the estimate for a quicker reaction to changes of the underlying distribution. The maximum achievable accuracy, in terms of the minimum Standard Error for a binomial proportion estimator implementing ageing, is depicted in Figure 15, p. 152.

Figure 15 depicts the theoretically achievable minimum Standard Error, which depends on the unknown, estimand parameter p. Extending on this, Figure 16, p. 153 shows the accuracy of Bayesian estimators for binomial proportions[26] in terms of the mean root-squared errors in a simulation setting, instantiated with a uniform prior and varying ageing factors a. In Figure 16a, mean root-squared errors are reported for a Monte-Carlo simulation of 10,000 runs of a stationary Bernoulli process with $p = 0.9$ and length 200. In Figure 16b, non-stationarity is assumed and the Bernoulli process generating the binary sequence changes its *probability of success*, p, twice. Initially, the *probability of success* is set to $p = 0.3$, which increases suddenly to $p = 0.6$ at time step 31, and further to $p = 0.9$ at time step 131.

26 See Chapter 3



(a) Stationary Bernoulli Process, $n = 200, p = 0.9$

(b) Non-Stationary Bernoulli Process, $p = 0.3$ (time 1-30), $p = 0.6$ (time 31-130),$p = 0.9$ (time 131-200)

Figure 16: Average accuracy, in terms of Root Mean Squared Error (Monte-Carlo simulation, 10,000 repeats).

Under stationary conditions, it is clearly visible that the estimator with no ageing outperforms those that implement ageing, in terms of the root-squared error. For a longer sample length ($n = 2,000$) and a more comprehensive set of ageing factors, see also Appendix E, Figure 34, p. 243.

Under non-stationary conditions, the advantages of implementing ageing are pronounced. For the given scenario, the basic estimator without ageing still exhibits the lowest overall root-squared error scores at times 30 and 130; however, it is very slow to react to changes in the Bernoulli process. In fact, it only narrowly outperforms the estimator implementing a conservative ageing factor of $a = 0.9$ from time step14 to 30, and again briefly from time step 121 to 130. Here, the application of a conservative ageing factor of $a = 0.9$ yields a significant improvement in the overall predictive performance of the estimator by maintaining a margin of error roughly comparable to the base estimator and providing significantly improved responsiveness to changes in the Bernoulli process.

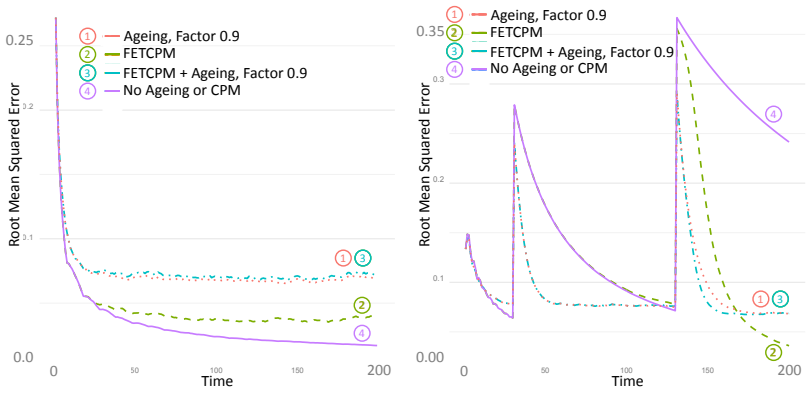
Change point detection can be applied to trustworthiness estimators, irrespective of whether or not ageing has been introduced. Change point detection increases the responsiveness of the estimator to change, without compromising its maximum achievable accuracy. Figure 17, p. 154, shows the effect of the FETCPM change detector on estimator accuracy and responsiveness for the same scenario as Figure 16b.

In Figure 17b, non-stationarity with two change points is assumed. Here, the change point detector improves the responsiveness of the base estimator without ageing considerably after time step 131, at which a change point occurred and was detected. The earlier change point at time step 31 was not reliably detected. However, the perfor-

mance of the estimator was not impaired by the *Type II* error. In fact, incorrectly failing to reject the null-hypothesis of the *FET* – h_0 ="*No change point has occurred*" – simply means that the base estimator is maintained in its original state. False positive change point detection, that is, *Type I* error, has an effect on estimator performance. The decreased performance of the estimator, caused by early detection of the change point at time step 131, manifests itself in a slight increase in the root-squared error just prior to time step 131.

Change point detection when combined with conservatively aged estimators, for instance, at $a = 0.9$, yields a further improvement in the responsiveness to change of the aged estimator. This increase in responsiveness to change is visible in Figure 17, albeit less pronounced than for the non-aged estimator.

For the stationary scenario, Figure 17a, the *Type I* error manifests itself in decreased performance compared to the base case of no ageing or change point detection. However, performance under change point detection is still superior to that under ageing for an ageing factor of $a = 0.9$.



(a) Stationary Bernoulli Process, $n = 200, p = 0.9$

(b) Non-Stationary Bernoulli Process, $p = 0.3$ (time 1-30), $p = 0.6$ (time 31-130),$p = 0.9$ (time 131-200)

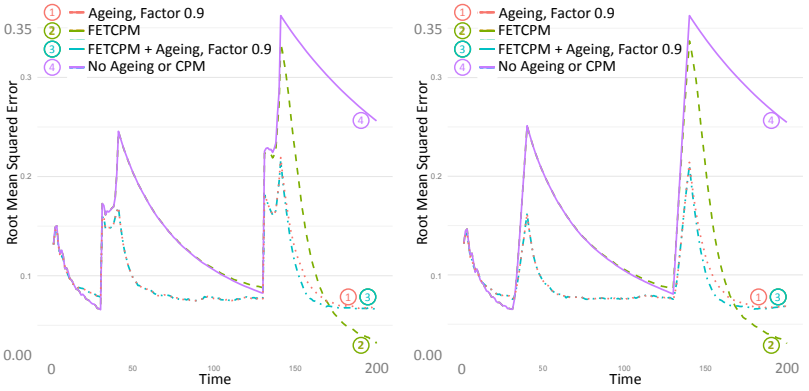
Figure 17: Average accuracy, in terms of Root Mean Squared Error (Monte-Carlo simulation, 10,000 repeats), with change point detection.

Figure 18, p. 155, shows the behaviour of ageing and change point detection under gradual and incremental change (see, Figure 14, specifically 14c and 14b, p. 139). Qualitatively, the behaviour remains unchanged in comparison to Figure 17b, p. 154.

Three different kinds of changes have been addressed in the simulation: sudden (Figure 14a), incremental (Figure 14b) and gradual (Figure 14c) change. Reoccurring change has not been addressed and is considered a special case of repeated sudden change. The responsiveness of the detector incremental and gradual change was affected by the length of the transition period inherent to inherent and grad-

ual change. However, this period also impacts the efficiency of ageing approaches as well. The relative difference between change point detection and ageing generally remains similar under different types of change.

Although not shown in diagrams, the detection time is dependent on the extent of the change. Smaller changes in the underlying parameter of the Bernoulli distribution take longer to manifest in the data in a significant way, consequently delaying the detection. However, small changes also have small effects on the point estimate, which in turn eases the need for a detection of changes. General results on the expected delays can be found in [178].



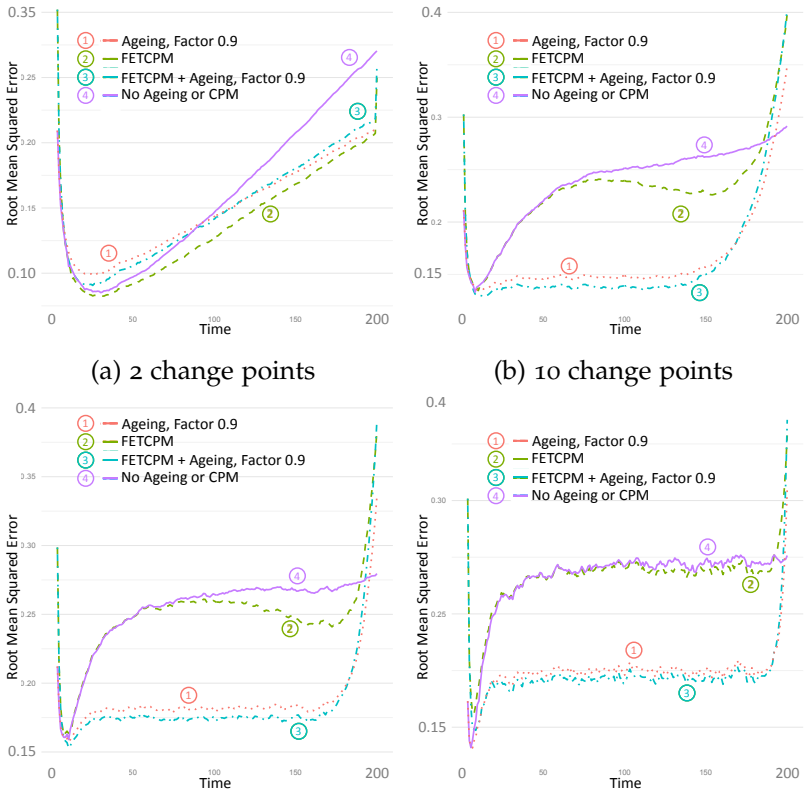
(a) Non-Stationary Bernoulli Process, $p = 0.3$ (time 1-30), $p = 0.6$ (time 41-130),$p = 0.9$ (time 141-200), gradual change at 31-40 and 131-140.

(b) Non-Stationary Bernoulli Process, $p = 0.3$ (time 1-30), $p = 0.6$ (time 41-130),$p = 0.9$ (time 141-200), incremental change at 31-40 and 131-140.

Figure 18: Average accuracy, in terms of Root Mean Squared Error (Monte-Carlo simulation, 10,000 repeats), with change point detection, under gradual and incremental change.

In order to investigate the effect of random change point positions on the effectiveness of ageing and change point detection, a Monte-Carlo style simulation randomising the position of the change points and the magnitude of the change for a given number of change points over a 200 time step sequence was simulated. This allows drawing further conclusions on the efficacy of both ageing and change point detection. Overall, conservative ageing (e.g., with factor $a = 0.9$) provides a good trade-off between accuracy and responsiveness to changes (Figure 19, p. 156).

In scenarios with a relatively low number of uniformly distributed change points, change point detection offers considerably improved performance when combined with a conservatively aged estimator. Once the number of randomly distributed change points increases, applying change point detection causes increases in the *Type I* error of the *FET*, leading to a *decreasing advantage* in performance of the



(a) 2 change points

(b) 10 change points

(c) 20 change points

(d) 50 change points

Figure 19: Average accuracy, in terms of Root Mean Squared Error; uniform random changes in p and location of change points (Monte-Carlo simulation of a non-stationary Bernoulli Process, $n = 200$, $10,000$ repeats).

FETCPM augmented estimator. As can be seen in Figure 19, p. 156, the performance of the estimator that is augmented with the FETCPM change detector is at worst on the same level as the non-augmented version. The uptick in the root-squared error at the latter time steps (> 150) is an artefact of the fixed-length sequence.

While the simulation settings chosen in Figure 19 offers an indication that the combination of change point detection and ageing has a positive effects. In order to gain a a more general insight, possibly going beyond the application of change point models in trustworthiness assessment, a closer examination of the benefits of combining change point detection and multiplicative ageing under various other scenarios might prove worthwhile. However, this is considered beyond the scope of this thesis.

4.4.7 *Section Summary*

The preceding section introduces non-stationarity[27] as a challenge to correct trustworthiness estimation. Instead of stationarity in general, only local stationarity is assumed, thereby introducing the concept of *change points* at which trustee behaviour changes.

By applying state-of-the-art change point detection methods [178] to trustworthiness estimation tasks, these behavioural changes can be detected and the point in time at which they occur can be estimated. This permits a potentially more accurate estimation than would be possible with methods based on multiplicative ageing, as they exhibit a lower bound on the accuracy that is theoretically possible to achieve – and hence on the achievable maximum certainty score. Additionally, when combined with ageing, change point detection increases the responsiveness to change over using either change point detection or ageing by itself.

The selected change point detector [178] is centred around the *FET* statistic that has already been introduced and used in previous sections. It is applicable to both binomial and mulitnomial trustworthiness estimation tasks; the latter is achieved by exploiting the fact that the marginal distributions of the (mulinomial) Dirichlet follow (binomial) Beta distributions.

4.5 CHAPTER SUMMARY

In this chapter, the core of the trust model introduced in Chapter 3 has been augmented with further methods for facilitating trust propagation and accounting for concept drift in the behaviour of the trustees. All of these augmentations are applicable, by design, to both the binomial case and the multinomial case of trustworthiness estimation.

Specifically, the methods introduced and discussed in this chapter deal with determining recommender trustworthiness (Section 4.2, p. 100) and the combination of opinions (Section 4.3, p. 112), both of which are necessities for robust trust propagation. Additionally, non-stationarity in the data generating process, i.e., potentially changing and dynamic trustee behaviour, is introduced to the model assumptions (Section 4.5, p. 160).

The methods for combining opinions, as presented in the preceding chapter, consist of discounting, consensus and fusion operations. The operations for discounting and consensus were modified to fit the extended version of the *CertainTrust* model introduced in this thesis, *Multinomial CertainTrust*. This represents a necessary step in providing a comprehensive trust model that includes capabilities for trust propagation. Additionally, the fusion operation, a method for combining opinions for which the assumption of independence does not

27 In related work on trust models, this is sometimes referred to as *dynamicity*.

hold, has been adapted for use in *Multinomial CertainTrust*. The original version, which is essentially an averaging operation present in both *Subjective Logic* [104] and *CertainLogic* [77, 175], has been adapted to *Multinomial CertainTrust*. Weighted and conflict-aware extensions, first published in [77], have been presented and considerably extended to the multinomial case, including a novel method for computing the degree of conflict leveraging the *FET*. Together with the novel *FET*-based method for determining recommender trustworthiness, presented in Section 4.2, a comprehensive multinomial trust model with trust propagation capabilities is enabled.

Finally, dynamicity, in the form of non-stationary behaviour by a trustee, is addressed and handled by applying state-of-the-art change point detection to trustworthiness estimation. Compared to multiplicative ageing, the application of change point detection method does not affect the achievable accuracy of the trust estimator. When used in conjunction with ageing, change point detection improves the responsiveness of the estimator to behavioural change over either individual method.

Specific contributions in this chapter include:

- A novel estimation method for *recommender trustworthiness estimation* was introduced (Section 4.2) that compares favourably to the related work (Section 4.3.6). This *FET*-based recommender trustworthiness estimation method can be applied to multinomial opinions, as opposed to the methods from the related work, which are applicable to binomial opinions only.

- Operations for trust propagation have been adapted or newly introduced for the use with *Mutlinomial CertainTrust*, specifically:
  - *Discounting* was adapted to *Mutlinomial CertainTrust* opinions,
  - *Consenus* was adapted to *Mutlinomial CertainTrust* opinions,
  - *Average Fusion* was adapted to *Mutlinomial CertainTrust* opinions,
  - *Weigthed and Conflict-aware Fusion* were newly introduced and expanded from [77], including a novel way of computing the degree of conflict between opinions.

- *Change point detection* was introduced into trustworthiness estimation in order to improve the trust model's responsiveness to dynamicity, expressed as non-stationarity.

The contributions in this chapter extend the trust estimation techniques presented in the previous chapter to the combination of opinions . Thereby, they enable accurate trust computation in scenarios where trust opinions from several sources have to be combined. Statistically well-founded approaches have been introduced, making the

combination of opinions sounder, particularly in multinomial models. This provides more exact trust estimates that are more readily interpretable from an estimation-theoretic point of view, allowing for accurate reputation systems, for instance, in industrial applications.

Overall, this chapter has provided the extensions necessary to make the trustworthiness prediction model described in Chapter 3 a comprehensive trust model with trust propagation capabilities.

5

EXTENSIONS FOR PRACTICAL TRUSTWORTHINESS ESTIMATION

In the previous Chapters 3 and 4, a considerably extended and mathematically much more rigid version of the *CertainTrust* model was introduced. These extensions provided trustworthiness assessment for multinomial responses, and included novel methods for the estimation of the certainty parameter, the estimation of recommender trustworthiness, as well as the detection of changing trustee behaviour, among others. *CertainTrust*, in both its basic [173] and extended versions, is fundamentally relying on Bayesian update learning in order to derive a trust estimate. Along with other, similar trust models, it is an evolution of basic feedback-based estimation techniques applied to trust, for instance work by Jøsang [103, 107, 108], Buchegger [28], Mui [151] or Teacy [189].

The Bayesian learning approach presented in Chapters 3 and 4 relies on a feedback-based data generating process, in which feedback, in the form of *experiences*, has to be collected by each truster *on each trustee*. While this burden is eased by the introduction of recommendations, no other indicators of trustworthiness, other than past experiences with a *specific trustee* are used for computing the trustworthiness of that truster. However, even in online environments, where traditional cues for trusting are not generally applicable anymore [141], information beyond experiences exists that might prove useful in trustworthiness estimation.

Additional information that can be harnessed for trustworthiness estimation may be derived from associations that a trustee maintains and that can be used to transfer trust, from demonstrating an investment by the trustee that would be lost in case of trustee deception, or from identifying attributes that trustworthy trustees typically exhibit. By finding and leveraging such information that is *indicative* of trustworthy behaviour, the estimation of the trustworthiness of trustees that have not been encountered by a specific truster before can be improved. Although prior experience may not exist or be sparse between a truster and a specific trustee, practical extensions to feedback-based trust models can be designed to provide those trusters with initial opinions of these unknown trustees.

In this chapter, work already published as two conference papers ([84] and [85]) is presented:

- Sascha Hauke, Florian Folk, Sheikh Mahbub Habib, and Max Mühlhäuser. Integrating Indicators of Trustworthiness into Reputation-

based Trust Models. In *Proceedings of the 6th IFIP WG11.11 International Conference, IFIPTM 2012*, 2012, [84] and

- Sascha Hauke, Sebastian Biedermann, Max Mühlhäuser, and Dominik Heider. On the Application of Supervised Machine Learning to Trustworthiness Assessment. In *Proceedings of the 12th IEEE International Conference on Trust, Security and Privacy in Computing and Communications, IEEE TrustCom-13*, 2013, [85].

In the first part of this chapter, Section 5.1, *CertainTrust* is extended by integrating so-called indicators of trustworthiness into the trustworthiness assessment and into the trust-based decision process. Three specific cases of indicators are presented: *insurance* of interactions, *certification* of trustees, and *coalitions* among trustees. These indicators are modelled as extension to *CertainTrust* and evaluated in an agent-based simulation. In the second part of this chapter, Section 5.2, the application of supervised machine learning techniques to trustworthiness estimation is addressed. The practicability of this approach is evaluated on a real-world data set of hotel features and ratings, and a way of integrating supervised estimates with *CertainTrust* is proposed.

5.1 HARDCODING INDICATORS OF TRUSTWORTHINESS

In this section, the integration of a select set of so-called *indicators of trustworthiness* into the *CertainTrust* model will be shown for three specific indicators, based on the concepts of insuring, certifying and coalition forming. In general, an indicator of trustworthiness will, in the following, be understood as per the following definition:

Definition 37 (Indicator of Trustworthiness)**.** An *indicator of trustworthiness* is any observable feature or combination of features exhibited by a trustee P that allows a truster A to infer the trustworthiness of trustee P.

Clearly, Definition 37 is general enough that it covers both samples of personal past experiences, $X = \{x_1, x_2, \ldots, x_n\}$, and opinions, $o = (r, s)$, as used for Bayesian trustworthiness estimation in the previous chapters. Thus, both a truster's own past experience with a trustee and recommendations regarding that trustee are indicators of trustworthiness. The core assumption of Bayesian computational trust models, of course, is that past performance indicates future performance.

Additionally however, it would be advantageous to identify other indicators of trustworthiness that provide a general link between an exhibited feature and a degree of trustworthiness. One way of doing so is to identify some specific identifiers 'by hand' and modify the prediction model accordingly. This approach will be introduced

here, given three specific indicators of trustworthiness that may be observed, for instance, in (electronic) commerce.

5.1.1 *Approach and Methods*

So far within this thesis, the focus has been on the relevant estimation of a trustee's trustworthiness and the concordant certainty. For this, Gambetta's definition of trust as a subjective probability [64] provided a useful foundation, that has been adapted into Definition 2 in Chapter **??**. Within the scope of this section, however, the trust concept has to be subdivided according to [112], by differentiating *reliability trust* (see Definition 3) from *decision trust* (see Definition 4).

Recall that the definition of reliability trust covers its use as an estimator of trustworthiness, i.e., the notion that trust is a subjective probability (see Chapter 2.1.2, pp. 20). Thus, the estimates computed by the *CertainTrust* trust model are representative of such an estimator. When having to make a decision, however, further considerations are involved, beyond the supposed reliability expressed by a trustworthiness estimate. This is reflected in the definition of decision trust [112], Definition **??**, which defines it as '*the extent to which a given party is willing to depend*'.

In a manner of speaking, reliability trust can be said to *inform* decision trust. However, *risk, gain, loss* and *reliance* [162] are also contributing to the decision-making process. Consequently, decision trust is generally modelled using expected utility theory [109, 131]. The probabilities, denoted as p, used in the computation of the expected utility will be derived from reliability trust estimates. That is, the values of various instances of p, e.g., used in equations 22 and 26, are trustworthiness estimates of *CertainTrust*. Expected utility, as a measure of decision trust, will be computed for the three extensions presented in the following: *insurance* (Section 5.1.2), *certification* (Section 5.1.3), and *coalitions* (Section 5.1.4).

The expected utility used as a basis for determining decision trust is defined thusly:

Definition 38 (Expected Utility (Binomial Case))**.** Let G be a benefit expected from an interaction, i.e., the positive gain, and L the corresponding loss, or negative gain. Furthermore, let $p \in [0, 1]$ be the probability of a beneficial outcome. Then, the expected utility EU of an interaction can be defined as [109, 131]:

$$EU := p \cdot G - (1 - p) \cdot L \tag{20}$$

In case the outcomes of an interaction are not binary but rather categorical, so that there are $m \in \mathbb{N}, m > 2$ different outcomes, Definition 38 is extended according to the following definition:

Definition 39 (Expected Utility (Multinomial Case))**.** Let $\mathbf{G} = (G_1, G_2, \ldots, G_m)$ be an m-dimensional, $m > 2$, vector of values signifying the gain from an interaction outcome of categories $1, 2, \ldots, m$, so that G_i represents the positive or negative gain resulting from an outcome in category i. Without loss of generality, let $G_1 \geqslant G_2 \geqslant \ldots \geqslant G_m$. Furthermore, let $\mathbf{p} = (p_1, p_2, \ldots, p_m)$ be a categorical probability distribution, with p_i representing the probability that an outcome of category $i \in \{1, 2, \ldots, m\}$ occurs. Let outcomes $1, 2, \ldots, m$ be exclusive and exhaustive, and therefore $\sum_{i=1}^{m} p_i = 1$. Then the expected utility EU of an interaction can be defined as:

$$EU := \sum_{i=1}^{m} p_i \cdot G_i \tag{21}$$

In the following, we will assume that the values of G and L, or $\mathbf{G} = (G_1, G_2, \ldots, G_m)$ for the multinomial case, are given constants and determined by the type of interaction in which trustee and truster are engaging. The actual value for p, or $\mathbf{p} = (p_1, p_2, \ldots, p_m)$, is unknown and is estimated by using *CertainTrust*. Because single value estimates are required, p will be given by a *CertainTrust* expectation value (Section 3.1.7, p. 69) for the binomial case, and $\mathbf{p} = (p_1, p_2, \ldots, p_m)$ by a m-dimensional *Multinomial CertainTrust* expectation value (Section 3.2.7, p. 88).

While absolutely trusted third parties are popular for externalising trust concerns in IT-security scenarios, it is generally not clear as to why these third parties that act *warrantors of trustworthiness* should deserve the status of being wholly and absolutely trusted. In the framework of probabilistic trust, such as the one assumed in computational trust models such as *CertainTrust*, however, absolute trust is nigh-impossible to achieve but, at the same time, also not a necessity. If trusted third parties are considered self-interested actors in their own right, they can still function as warrantors; in fact, such warrantors can be considered service providers themselves, providing *trust-building services*. Prime examples of trust-building services are, for instance, the issuing of certificates or the provisioning of insurance. At the core of these services –generally – rests the delegation of the warrantor's trustworthiness to the trustee. Therefore, trust-building services are only valuable – to both the trustee using them and the warrantor offering them – if the warrantor that provides them is (highly) trusted and persistent.

In the following, three types of trust-building services are introduced, one (*insurance*) primarily affecting decision trust. The other two (*certification* and *coalition forming*) affect the computation of reliability trust, which influences the probabilities used in decision making.

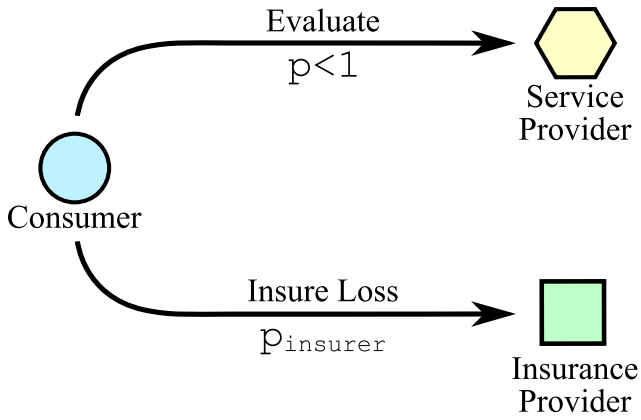


Figure 20: Trust delegation with insurance.

5.1.2 *Reliance through Insurance*

The insurance service incorporates three entities: The truster, a *consumer* trying to identify the most appropriate service provider to select, the trustee, *service provider* under evaluation, and the warrantor, an *insurance provider* insuring the transaction if the consumer decides to interact with the service provider. The relations between the entities are outlined in figure 20.

An important concept in terms of insurance is the concept of *reliance*. Following [162], reliance is defined in the following Definition 40:

Definition 40 (Reliance)**.** Reliance is the act of trusting a third party or institution to prevent the truster from incurring a permanent loss. This can be achieved by direct control of the third party over the trustee, by means of coercing the trustee to act trustworthy or by affording redress in case of untrustworthy behaviour of the trustee.

Insurance provides reliance, and thus affects decision trust, by reducing the risk of asset loss attendant with an interaction. It therefore should contribute to *"[...] a feeling of relative security [...]"* (Definition 4). In the following, it will be assumed that the fulfilment of the trustee's obligation towards the truster, thus, whether or not the trustee has acted in a trustworthy manner, can be objectively determined. Specifically, trustee performance is seen as an objective measure that both the insurer and the truster agree upon.

In terms of expected utility, the insurance scenario can be formalised as follows: Let $p_{trustee}$ be the probability of a successful interaction with a trustee service provider, and $p_{insurer}$ the probability of a successful interaction with an insurance provider that vouches for or guarantees the interaction between the truster consumer and trustee the service provider. Furthermore, let the cost, or negative gain, the truster consumer experiences in case of an unsuccessful interaction

with the trustee service provider, be denoted $L_{trustee}$. Analogously, $L^{fix}_{insurer}$ is the cost (if any) of the insurance contract to the consumer. Additionally, $L^{var}_{insurer}$ indicates the expenses incurred by the consumer when making an insurance claim against a failed interaction. In this case, the expected utility of the interaction for the consumer is:

$$\begin{aligned} EU := \; & p_{trustee} \cdot G \\ & - (1 - p_{trustee}) \cdot (1 - p_{insurer}) \cdot (L_{trustee} + L^{var}_{insurer}) \\ & - (1 - p_{trustee}) \cdot (p_{insurer}) \cdot L^{var}_{insurer} \\ & - L^{fix}_{insurer} \end{aligned} \tag{22}$$

The expected utility in the insurance case is the probability of the trustee to act in a trustworthy manner, times the expected gain from the interaction. From this, all the cases, in which the trustee fails have to be subtracted; in case the insurer also fails to meet its obligations, the truster incurs a loss of magnitude $L_{trustee} + L^{var}_{insurer}$, in case the trustee fails, but the insurer redresses the truster, this loss is reduced to $L^{var}_{insurer}$. Additionally, any fixed cost arising from the insurance options itself will also have to be deducted.

Adapting the insurance scenario to the multinomial case requires a stringent definition of the insurer's actions in case an interaction falls into any of the categories not considered fully satisfactory. This requires a negotiation between truster and insurer, as to what the compensation is for each outcome category. In the following, it will be assumed that an agreement between the truster and the insurer has been reached regarding what outcomes will be compensated through insurance.Therefore, a new term R_i, denoting the redress provided by the insurer in case of an outcome of category $i \in \{1, 2, \ldots, m\}$ is introduced. Additionally, the function $\mathbb{I}_I(i)$ is the indicator function over the set of categories for which redress has been agreed upon by the truster and the insurer. In terms of the multinomial expected utility formulation of Definition 21, p. 164, the insurance case can be formalised as:

$$\begin{aligned} EU := \; & \sum_{i=1}^{m} p_{trustee,i} \cdot G_i \\ & + p_{insurer} \cdot \sum_{i=1}^{m} (p_{trustee,i} \cdot \mathbb{I}_I(i) \cdot (R_i - L^{var}_{insurer})) \\ & - (1 - p_{insurer}) \cdot \sum_{i=1}^{m} (p_{trustee,i} \cdot \mathbb{I}_I(i) \cdot L^{var}_{insurer}) \\ & - L^{fix}_{insurer} \end{aligned} \tag{23}$$

In the formalisation of the multinomial case, G_i denotes both positive, as well as negative gain. The indicator function $\mathbb{I}_I(i)$ is introduced to account for the fact that only in those cases in which redress

for an outcome of category i has been agreed upon between truster and insurer, the truster will incur the cost of claiming the insurance. Otherwise, the Equations 22 and 23 are equivalent.

Interaction		Update	
Trustee	Insurer	Trustee	Insurer
success	–	positive	–
failure	success	negative	positive
failure	failure	negative	negative

Table 8: Trust Updates with Insurance.

After an insured interaction between a truster consumer and the selected trustee service provider took place, the truster updates its trust values according to Table 8. In case the interaction with the trustee succeeded, additional positive evidence regarding the trustee is created. In this successful case, action from the insurer is not demanded and no further evidence regarding the insurer is collected. However, if the interaction with the selected candidate fails, there are two possible cases. If the insurer is called upon and reimburses $L_{candidate}$ to the consumer, therefore compensating the negative gain for the consumer, new positive evidence for the insurer is collected. If the insurer fails in compensating the negative gain, new negative evidence regarding the insurer is collected, e.g., by increasing the value of $s_{insurer}$. In both cases, new negative evidence regarding the selected provider is created analogously.

The insurance case considerably reduces the loss in case of trustee deception, thereby boosting the expected utility of an interaction with an insured trustee. This affects decision trust positively. A discussion within an illustrative running example and agent-based simulation results are presented in Section 5.1.5.

5.1.3 *Assessing Reliability through Certification*

Similar to the insurance case above, the certification procedure consists of three interacting entities. The truster consumer is evaluating a a trustee service provider for selection. This service provider is certified by a warrantor, in the form of a certification provider, whom the consumer has prior knowledge about but does not interact with directly (see figure 21).

It is assumed that a certification provider certifies service quality for an entire service or a service component. Certification of partial aspects of a service (component) can be combined into an overall rating, for instance by using the propositional logic operators of *CertainLogic* [78, 175]. Formally, a certification describes a specific minimum

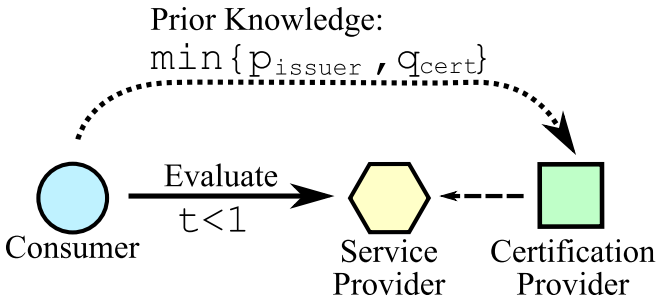


Figure 21: Trust delegation with certification.

level of quality as $q_{cert} \in [0, 1]$ that a certification provider awards to the certified party, ideally after completing an audit.

This kind of limited trust delegation, employing a probabilistic certificate value and a certification provider that is not necessarily a completely trusted third party, influences the reliability trust the truster places in the trustee. In particular, in order to preserve the importance of direct experience over other kinds of information, we propose to include certification information in the initial expectation value f of CertainTrust. In its simplest form, it is defined it as follows:

$$\begin{aligned} p_{issuer} &= E(t_{issuer}, c_{issuer}) \\ &= c_{issuer} \cdot t_{issuer} + (1 - c_{issuer}) \cdot f \\ f_{cert} &= max(f, min(p_{issuer}, q_{cert})) \end{aligned} \tag{24}$$

$$E^{cert}(t_{trustee}, c_{trustee}) = c_{trustee} \cdot t_{trustee} + (1 - c_{trustee}) \cdot f_{cert}$$

The variables c_{issuer}, t_{issuer}, $c_{trustee}$, $t_{trustee}$, and f are derived using *CertainTrust*.

In case the initial trust value f is to be computed from a number of different sources, for instance by combining different opinions using *CertainLogic* [78, 175], it is useful to extend Equation 24 so as to represent the certificate as a *CertainTrust* opinion. For this, let the certificate be represented by *CertainTrust* opinion $\omega_q = (q_{cert}, c_q)$, and $\omega_{issuer}^{truster} = (t_{issuer}, c_{issuer})$ be the opinion that the truster has on the issuer of the certificate. $\omega_{issuer}^{truster}$ is the opinion of the truster on the certificate issuer. Implicitly, it is assumed here that certificate issuers are relatively well-known entities that have a reputation, thereby allowing the truster to derive an opinion on the certificate issuer from its own past experiences or recommendations.

Using the *CertainTrust* robust discounting operation, ω_{cert} results from ω_q discounted by the quality of the issuer, as evidenced by $\omega_{issuer}^{truster}$. For robust discounting with *CertainTrust*, see Chapter 4.3.5, p. 126. The resulting opinion on the trustworthiness of the trustee, ω_{cert}, is contingent on whether or not the certificate issuer is considered trustworthy *and* the quality level reported in the certificate.

A number of assumptions are helpful for this. First, that the certificate issuer only issues certificates if it is reasonably certain that the certified party is – at least – trustworthy at the probabilistic value that was given in the certificate; that is, the certificate issuer is confident in its certification procedures and that they were executed appropriately. Therefore, the certainty value in opinion $\omega_q = (q_{cert}, c_q)$, c_q, should be 1 or close to 1. Second, the certificate issuer itself should enjoy a good reputability.In particular, trust factors such as institutional or system trust [142] can be placed in certificate issuers that are institutions or act within an institutional framework. High trust in the institutions and systemic guarantees within a society transfers to the reputability and perceived trustworthiness of a certificate issuing institution. ISO[1], as a standardisation body, is a prominent example.

The resulting *CertainTrust* opinion ω_{cert} can be fused with other opinions $\omega_1, \ldots, \omega_n$ using the fusion operations described in Chapter 4.3.3, p. 114, to be used as an input for a parameterisation of the *CertainTrust* initial trust value, f, as an informative prior (see Chapter 3.1.6, p. 66). Assuming that ω_{cert} is not fused with any other opinions representing prior information, Equation 24 is modified to become:

$$f_{cert} = E(\omega_{cert}) \tag{25}$$

$$E^{cert}(t_{trustee}, c_{trustee}) = c_{trustee} \cdot t_{trustee} + (1 - c_{trustee}) \cdot f_{cert}$$

The reliability trust score $E^{cert}(t_{trustee}, c_{trustee})$, either according to Equation 24 or Equation 25, informs expected utility computation for decision trust. $p^{cert}_{trustee} = E^{cert}(t_{trustee}, c_{trustee})$ is the probability estimate for a successful interaction with a candidate service provider, given a certification from a certification provider.

Then, the expected utility of the interaction between a consumer and a certified service provider can simply be described as:

$$EU := p^{cert}_{trustee} \cdot G - (1 - p^{cert}_{trustee}) \cdot L \tag{26}$$

For the multinomial case, the formulation is analogously defined. Instead of providing a single-value quality estimate q_{cert}, the certificate issuer reports an m-dimensional multinomial vector, $\mathbf{q} = \{q_1, q_2, \ldots, q_m\}$ and concordant certainty values in an m-dimensional *Multinomial CertainTrust* opinion. Applying the extended multinomial expectation value computation described in Chapter 3.2.7, 88, this yields an m-dimensional vector of expectation values, which can be used to derive an informed multinomial prior:

$$f^m_{cert} = E(\omega^m_{cert}) \tag{27}$$

1 http://www.iso.org/

The m-dimensional vector is then used as the initial trust values for the, also m-dimensional, *Multinomial CertainTrust* opinion of the truster on the trustee. Applying the extended multinomial expectation value computation described in Chapter 3.2.7, 88, on this opinion gives an m-dimensional vector of probability estimates that can be used to inform the multidimensional decision trust. This is done by instantiating these probability estimates into Equation 21, p. 164.

Trust updates after an interaction are done for both the trustee and the certificate issuer. Updates for the former are handled according to the standard *CertainTrust* update strategy, while those for the latter use the update mechanism for determining recommender trustworthiness (see Chapter 4.2, p. 100).

5.1.4 *Joint Reliability through Coalitions*

Another way for trustees, for instance, service providers, to represent their trustworthiness is the formation of coalitions with other trustees. The motivation behind the introduction of this mechanism is the underlying assumption that a mutual association with another trustworthy entity serves as an indicator of trustworthiness. Lack of experience with one service provider, i.e., the trustee, can thus be compensated by the consumer, i.e., the truster, via the delegation of trust from associated service providers, i.e., the truster's associates, that might be known to the consumer.

While a coalition is different from an upfront monetary investment that insurance or certification represent, it is unlikely that established providers form coalitions with service providers that are unknown to them. A coalition requires its participants to perform well in order to *belong* to a select group of reliable members. This group is based on mutual benefit, and the collateral at its core is social. Simple sybil attacks from malicious service providers that spawn many identities and create coalitions between them are unlikely – because they are ineffective: coalitions influence the probability of being selected by increasing the visibility of a trustee, e.g., a service provider. Being associated with a *well-known and trusted* party becomes an implicit certification. A mutual coalition of unknown service providers does not increase the visibility of the participants. More sophisticated collusion attacks are still possible. However, by using trust delegation in coalitions to inform the prior only, should limit the effects of deceptive behaviour. Additionally, in scenarios in which mutual cooperation between the members of a coalition is required, for instance, in service composites, trustworthiness delegation via coalitions seems warranted, as the component service providers actively need to rely on each other to successfully provision a service.

Assume a truster, e.g., a consumer, wishes to evaluate a candidate trustee, for instance, aservice provider. It lacks, however, past direct

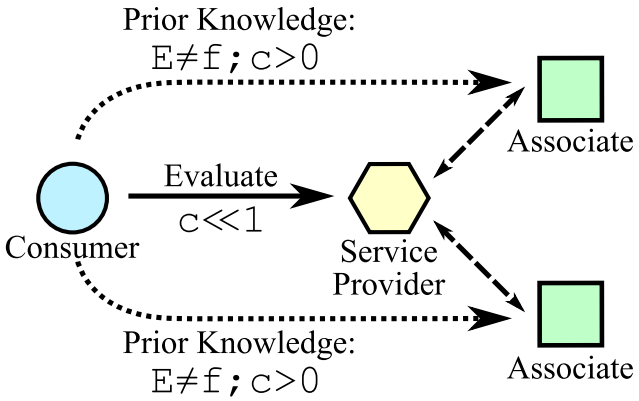


Figure 22: Trust delegation with associates.

experiences and recommendations to form a reliable opinion. This lack of knowledge might lead the consumer to choose another, better known service provider or forgo the interaction altogether. In order to alleviate the problem and be able to realise a profit from the interaction, it is in the trustee's best interest to increase the consumer's perception of its trustworthiness. To this end, the potential trustee service provider presents a list of other service providers it is associated with in a coalition to the consumer. As shown in figure 22, this is done under the expectation that the consumer has prior experiences with at least some of those. In this case, the experience the consumer has in the service provider's associates is partially transferred to the candidate.

REALISING MUTUAL COALITION In composed services, coalitions are already in place. By taking into account the nature of the cooperation of service composition sub-components and their respective providers, trust delegation through the proposed coalition mechanism is a feasible method of establishing trust. Whether or not such a delegation is appropriate is dependent on the direction of the trust delegation with regard to the order of the sub-components within the process, as well as on power symmetries and enforcement possibilities among the providers associated within a service composition. For instance, considering the illustrative running example presented at the beginning of Section 5.1, it can be argued that the credit card provider (i.e., visible component 1 in figure 24) is strongly connected to the grey box internal process. This is due to strong obligations and enforcement mechanisms (e.g., binding legal agreements and litigation possibilities) integrating the respective service providers.

If not explicitly cooperating in the service composition under evaluation, service providers that otherwise cooperate can enable coalition-based trust delegation through the following mechanism by advertising their cooperation to the customer. The customer, acting as truster, can consequently verify the coalitions and transfer trust accordingly.

Mutual coalitions are realised through the exchange and mutual acknowledgment of cooperation messages. A process for this is depicted in Figure 23.

1. Service provider A creates a message $m_{A,B} =< UID_A, UID_B, data >$ consisting of
   - a unique identifier representing provider A, e.g., an X.509 certificate
   - a unique identifier representing associate B, e.g., an X.509 certificate.
2. Service provider A forwards $m_{A,B}$ to service provider B.
3. B acknowledges its coalition with A by signing $m_{A,B}$.
4. B returns the signed cooperation message $\{m_{A,B}\}_{sigB}$.
5. A forwards its signed counterpart cooperation message $\{m_{A,B}\}_{sigA}$.

These cooperation messages can then be presented to potential consumers, in order to facilitate the coalition-based trust delegation.

6. A potential consumer C evaluating service provider A requests indicators of trustworthiness from A.
7. A supplies C with a list of cooperation messages.
8. C may validate the coalition between A and B by requesting B to verify the signed cooperation message $\{m_{A,B}\}_{sigB}$.
9. Service provider B, as an associate of A, either confirms or denies the coalition with A, in particular regarding both the validity of the signature and currentness of the coalition (A coalition is not current anymore if it existed in the past (when the messages were exchanged and signed) but has since been revoked by at least one of the parties).
10. The consumer C delegates the trustworthiness of B to A.

DELEGATING TRUST IN COALITIONS The delegation of trust from associates to the trustee functions similarly to the delegation via certificates. Instead of a certificate issuing warrantor, however, the associates' own trustworthiness serves as an implicit 'certification' score. Let $\omega_{A1}, \omega_{A2}, \ldots, \omega_{An}$ be the *CertainTrust* opinions the truster, e.g., a consumer C, has on associates $A_1, A_2, \ldots, A_n$ with which a trustee, e.g., a service provider, forms a coalition. Using the conflict-aware fusion operation described in Chapter 4.3.3, p. 114, the resulting fused is used as an input for a parameterisation of the *CertainTrust* initial trust value, f, as an informative prior (Chapter 3.1.6, p. 66). Additionally, let the fused opinion $\omega_{A1,\ldots,An}$ be discounted by a discounting

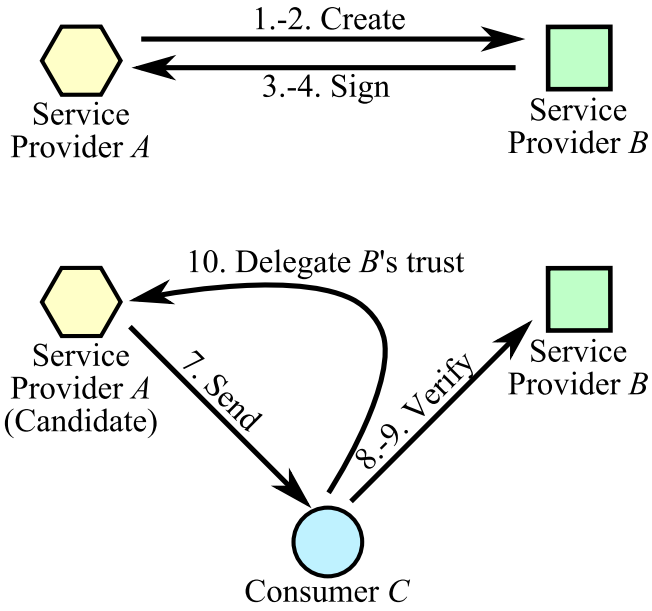


Figure 23: Coalition forming and verification of cooperation messages.

factor $\delta \in [0;1]$, based on the number associates in the coalition. For the discounting function, a linear function $f(n) : \mathbb{N}^+ \mapsto [0;1]$ is assumed. This function accounts for the number of associates that the truster considers necessary to accept the fused average opinion as representative.

The resulting Equation 28 is otherwise equivalent to Equation 25, p. 169.

$$f_{assoc} = E(\delta \cdot \omega_{A1,\ldots,An}) \tag{28}$$

$$E^{assoc}(t_{trustee}, c_{trustee}) =$$

$$c_{trustee} \cdot t_{trustee} + (1 - c_{trustee}) \cdot f_{assoc}$$

Accordingly, the expected utility is also equivalent to the certification case (Equation 26, p. 169):

$$EU := p^{assoc}_{trustee} \cdot G - (1 - p^{assoc}_{trustee}) \cdot L \tag{29}$$

The multinomial case is covered as in the previous example on certification. It is handled equivalently to Equation 27, p. 169.

The trust updates after an interaction can be found in Table 9: only new evidence for the selected service provider is collected regarding its performance. The selected provider alone is responsible for its performance as the only influence of the associates is the association itself. The future performance of the associates is independent from the selected provider. If the service provider and the associate are not part of same service composition, new evidence for the associates is

Interaction		Update	
Provider	Associates	Provider	Associates
success	–	positive	see text
failure	–	negative	see text

Table 9: Trust Updates with Coalitions.

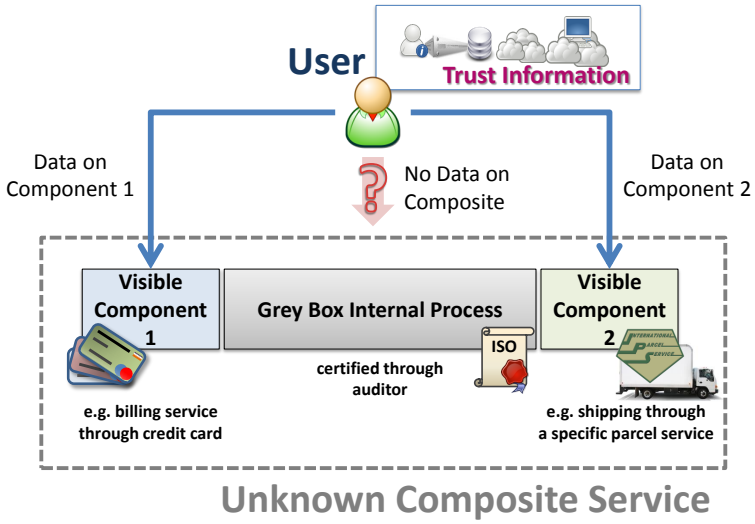


Figure 24: Running Example.

collected only in the context of their ability to reliably form associations. If they are, however, part of the same composite service (e.g., as in the running example presented at the beginning of Section 5.1), the reputation is updated for all service components.

5.1.5 *Evaluation*

5.1.5.1 *Running Example*

In order to illustrate the insurance, certification and coalitions, an e-commerce use case will be used. This running example, introduced for purely illustrative purposes, encompasses both virtual and concrete service components. Such services are commonly encountered in e-commerce scenarios, for instance when ordering physical goods online. Here, one portion of the service provisioning is digital, i.e., the ordering and payment processes, while the production and delivery of an actual, physical good are concrete, real-world processes.

For the running example, consider a customer trying to establish trust on a service. Furthermore, suppose that the (truster) customer does not have any prior experience with that particular service. It is therefore not immediately possible to derive the trustworthiness of the (trustee) service provider from direct experience. In order to

derive the reliability of the service, the conventional approach for feeback-based trust models (e.g., [93, 111, 173]) is to query trusted witnesses for information. However, even in the absence of reliable witnesses, it would be highly desirable to be able to at least roughly estimate the trustworthiness of the trustee service provider.

In the following, we will follow the intuition that in services, that involve both digital and real-world processes, such as online ordering and physical shipping of goods, service provisioning is generally not monolithic. Rather, the service provisioning processes can be subdivided into sub-components, some of which are visible to the truster customer and may be associated with distinct entities on which trust can be established individually.

Figure 24 outlines a general scenario in which a customer establishes trust on an unknown composite service. By necessity, several components of the service are visible to the customer, such as payment/billing and shipping agents used by the service provider. We assume that the billing process is handled through an intermediary, for instance a credit card company. For the core service provisioning process, we further assume that the composite service provider chooses not to reveal its internal processes to the customer directly. It may, however, use an external auditing and certification provider (e.g., ISO) to certify its internal processes. Thus, a certification is considered to be representative of the quality of the internal service provisioning process.

5.1.5.2 *Discussion within Running Example*

The running example presented above introduces a composite service, in which some service components/providers are visible to the users, while others are contained in a grey box internal process. We deem this running example to be typical of an online goods ordering process. The payment functionality for the service is provided through a credit card company, while the delivery is handled by an independent parcel service. The grey box process is certified by a certification provider.

INSURANCE THROUGH CREDIT CARD COMPANY It can reasonably be assumed that the credit card company is well-known to and trusted by the customer. This stems both from past experiences and (possibly more importantly) from strong contractual obligations between a customer and his credit card company. Similar obligations exist between the credit card company and the provider of the composite service. Thus, social and legal assurances are in place to enforce the dependability of the partners in this setting. Furthermore, because a large number of internet services use a small number of credit card companies, experience with the credit card provider generally increases more rapidly than experience with any particular com-

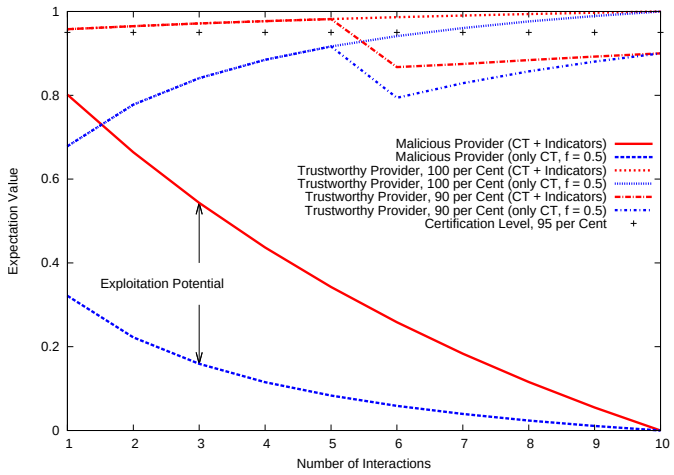


Figure 25: Reliability trust expectation, for $N = 10$ and $f = 0.5$.

posite service. Additionally, a credit card company within a service composition offers insurance services to its customers.

CERTIFICATION OF INTERNAL PROCESS Within the running example, the grey box internal process is certified by a certification provider (ideally following a thorough and transparent audit), for instance ISO (e.g., for quality management) or TRUSTe (for privacy, however cf. [50]). Certification providers are less strongly coupled with a service than the aforementioned credit card company. We assume that a limited number of certification providers is used by a considerable number of services, thus easing trust establishment on certification providers. Paying for a certification by a reputable certification provider indicates a service provider's initial commitment to remaining in a market (i.e., an incentive not to defect) [59].

Both insurance and certification depend heavily on reliance [162] on a third party. Trust in the insurance and certification providers to enforce user interests in case of service provider defection has to be established. If a certification provider is incapable or unwilling to enforce its certification rigorously, a certification can actually be interpreted as a sign of untrustworthiness [50]. It is therefore assumed that the user can reliably establish trust on insurance and certification providers using a trust model.

The shipping service represents the physical interface of the composite service to the customer. While the reliability of the shipping provider is essential to a successful overall service provisioning, it is not strongly coupled to the grey box internal process of the running example.

RELIABILITY TRUST COMPUTATION Modelling overall reliability trust in the unknown service composition requires combining the in-

formation on its components. Both a *coalition* and *certification* influence the computation of reliability trust.

- *coalition*: Due to the highly regulated relationship between the the credit card provider and the grey box internal component of the service composition, the providers of these two components are considered to be in a coalition. Therefore, the well-established trust the users has in its credit card provider is delegated to the internal component.
- *certification*: The internal grey box component of the composite service is assumed to be certified by an established certification provider, as per Figure 24.

As the all three service components are essential to the success of an interaction between customer and the service composition, the linkage between them can achieved by applying a belief logic-based conjunction operator, such as the *CertainLogic* AND operator ($\wedge_{CL}$) [175]. Including a certification provider to certify the grey box internal process (for which no prior experience is assumed to have been recorded), the overall computed reliability trust in the unknown composite thus becomes:

$$\begin{aligned} p_{composite} \approx\ & (t_{credit}, c_{credit}, f_{credit}) \\ & \wedge_{CL}(t_{grey}, c_{grey}, f_{grey}) \\ & \wedge_{CL}(t_{shipping}, c_{shipping}, f_{shipping}) \end{aligned} \tag{30}$$

with f_{grey}, the initial trust score for the grey box internal component, computed as the fusion between certification process and the delegated trust score from the credit card company.

Under a complete lack of information on *any* part of the composite service, i.e., the truster has no experience or recommendations on either the grey box internal component, the credit card company or the shipping service, the reliability trust value of the indicator-augmented trust computation corresponds to the *CertainTrust* value without indicators. The return value for $p_{composite}$ in this case is the truster's own initial expectation f.

INITIAL RELIABILITY TRUST SCORE AND ITS EVOLUTION Figure 25 shows the behaviour of the trustworthiness estimation using *CertainTrust* with and without indicators over 10 interactions (for $N = 10$ and $f = 0.5$). The trustworthiness of the credit card company and the certification provider were assumed to be high ($p = 0.95$) and known to the user at this level with certainty ($c = 1$). In this way, coalition and certification was essentially used to dynamically alter the initial trust in the unknown composite service, from $f = 0.5$ for the base *CertainTrust* case without indicators, to ≈ 0.92 (at certification quality, $q_{cert} = 0.95$).

While trustworthy service providers can thereby overcome cold start issues effectively, it theoretically offers malicious service providers a considerably bigger potential to exploit this positive reputation. This is depicted as *exploitation potential* in Figure 25.

The danger of malicious exploitation is counterbalanced by the fact that the increase of the initial trust expectation from 0.5 to ≈ 0.92, is not wholly arbitrary. Increasing the reliability trust in the unknown service was based on two criteria:

- First, that the certification provider (e.g., ISO) would audit the service provider and possible revoke the certification in case of a complaint against the service.
- Second, that the use of a credit card company affords strong reliance. Because the credit card company does not only stake its reputation, but also direct monetary values through an insurance service, it has a strong incentive to actually enforce the contractual obligations between itself and the core component of the unknown service composition (the grey box).

INCREASED DECISION TRUST THROUGH INSURANCE The reliance introduced through the credit card payment process does not only justify adjusting the initial expectation value of the reliability trust upwards, but also directly influences the customer's decision criterion, as per equation 22. This equation reflects the level of protection the credit card provider offers for an interaction with a possibly fraudulent service.

For the running example, we assume that the cost of the ordered good (this includes additional costs such as shipping & handling) is paid upfront through a credit card. This money is potentially lost in the interaction, it therefore represents $L_{candidate}$. The gain G is at least as high as $L_{candidate}$, otherwise it would be unreasonable to begin the transaction. The cost of claiming a credit card insurance is assumed to be negligible compared to the cost of the product, while the fixed costs of the insurance ($L^{fix}_{insurer}$) are covered via a surcharge on shipping and handling levied by the service provider.

Due to strong contractual agreements between the customer and the credit card company, the trustworthiness of the credit card provider (expressed as $p_{insurer}$) can be practically assured. Assuming that $L_{trustee} = G$ and $p_{insurer} \approx 1$, the decision criterion for the running example thus becomes:

$$EU := p_{composite} \cdot G - (1 - p_{composite}) \cdot (1 - p_{insurer}) \cdot G - L^{fix}_{insurer}$$

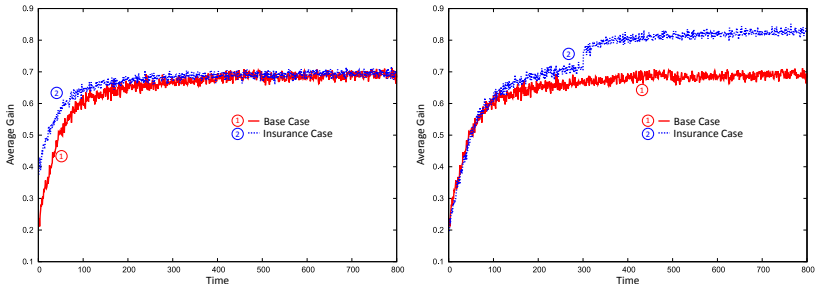
. For $p_{composite} \ll 1$, as would be the case when facing an unknown service, the expected utility is considerably higher for the insurance through credit card case than it would be without the insurance option. Thus, even under the risk of increasing the exploitation potential with regard to malicious service providers, reliance mechanisms,

such as insurance, provide a complementary measure helping the customer to feel safe.

5.1.5.3 *Simulation*

In order to show the feasibility of the proposed mechanisms in a qualitative way, each was implemented in an agent-based simulation framework, *MASCoT* [45].The *CertainTrust* trust model was used for evaluating providers, using *CertainTrust* parameters $f = 0.5, w = 1, N = 10$. The decision criterion used is expected utility, using the *softmax* approach [187] and a decaying temperature parameter to simulate exploitation-vs-exploration preferences. A consumer population of 250 agents was arrayed in a clustered social network (generated according to [100], with a clustering coefficient of 5), to serve as recommenders. The same basic configuration was used to test all mechanisms against a base case, which consisted of consumers solely using experience and recommendations to select providers. The market was started with 15 providers (5 with $0.8 < p_{trustee} \leqslant 0.95$, 5 with $0.5 < p_{trustee} \leqslant 0.8$ and 5 with $0 < p_{trustee} \leqslant 0.5$) and ran for 800 rounds. At round 300, a new provider with $p_{trustee} = 0.95$ is added, in order to test the market entry performance of the different mechanisms. The objective is for the consumers to select the best provider by learning the providers' trustworthiness.

During each time step, a randomly selected subset of consumers (20 per cent of the consumer population) evaluated the providers and interacted. Once a provider was selected by a consumer, and an interaction occurred, the consumer incurred either a gain of $+1$ utility or a loss of -1 utility, depending on whether or not the selected provider acted in a trustworthy manner. The average gain reported in Figures 26a, 26b and 27 is the average of the utility incurred by all those providers that interact during the current time step.



(a) Average gain with insurance (b) Average gain with certification

Figure 26: Agent-based simulation results for insurance and certification compared to base case.

INSURANCE As Figure 26a shows, over the entire simulation run, the performance of the insurance mechanism (measured as the av-

eraged gain over all consumers) approaches the base case. Significantly better performance, as determined by a Wilcoxon rank-sum test ($p < 0.01$)[203], was attained in the initial phase of the learning process, i.e., between time steps 0 and 250. In this early phase, the *softmax* algorithm exhibits a higher exploration rate, thus leading to a higher proportion of untrustworthy providers with $p_{trustee} \leqslant 0.5$. Losses incurred are compensated by insurance providers, represented as agents with $0.5 < p_{insurer} \leqslant 0.95$. The insurance providers were randomly assigned to the interactions. Parameters $L^{fix}_{insurer}$ and $L^{var}_{insurer}$, the cost of insuring a transaction, for the customer, are assumed to be covered by the trustee provider and thus negligible, i.e., 0.

Varying the reliability of the insurance providers, between a uniform $p_{insurer} = 0$ and $p_{insurer} = 1$ for all insurance providers, scaled the effectiveness of the mechanism. At $p_{insurer} = 0$, the insurance case showed no significant difference from the base case. At $p_{insurer} = 1$, the performance was marginally (statistically non-significant) better than the performance shown in Figure 26a.

CERTIFICATION The effects of certification (figure 26b) are complementary to the insurance case. While showing no improvement over the base case in the early rounds, it facilitates easier market entry for new providers with a high trustworthiness. 5 certification providers were introduced as separate entities. The certification providers are assumed to be honest and certify conservatively ($q_{cert} = p_{trustee} - 0.1$). Generally untrustworthy or marginally trustworthy providers ($p_{trustee} < 0.6$) were treated as though no certification was available, i.e., at $f = 0.5$. Certifier performance was learned using the *CertainTrust* model. The considerable improvement at time step 300 is caused by the addition of the new, trustworthy provider, which is selected based on its certification, despite *softmax* already being highly exploitative at this time step .

COALITIONS Coalitions outperform the base case (figure 27) after initial exploration significantly. This is caused by trustworthy providers dissolving coalitions with less trustworthy ones, leading to highly selected coalitions of good providers. For this simulation, coalitions are formed with up to 2 other providers. Each provider in a coalition operates non-competitively from its associates, i.e., the simulation was run with three different provider populations of 15 providers each. Only one such market is plotted, although the population of truster consumers was active in all three concurrently.

It is noticeable that the advantage of using coalitions, as opposed to the certification process, is more marked earlier on in the simulation (before time step 300, at which a new, highly trustworthy provider is introduced). This is caused by more abundant trust information

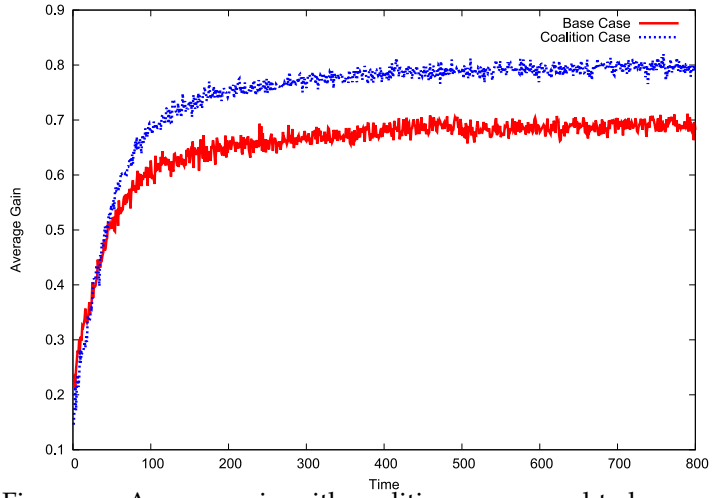


Figure 27: Average gain with coalitions compared to base case.

delegated from providers in other markets, in which the consumers acted as well. Thus, trust in the coalitions' associates is build faster than that in the certification provider in the previous case.

5.1.6 *Section Summary*

The preceding section introduced three mechanism (insurance, certification and coalitions) as indicators of trustworthiness for *CertainTrust* that influence the initial expectation (certification and coalitions) and the general perception of risk (insurance) of a customer towards a service. Each indicator has a distinct impact on the overall provider selection by consumer populations, allowing consumers to reduce their risk (insurance) and providers to represent their capabilities (certification and coalitions). By investing resources and staking reputation, service providers represent their commitment to a market, easing the service selection problem for the consumers.

In an exemplary simulation setting, the effect that the different indicators can have in a service selection scenario have been demonstrated. Statistically significant improvement over the base case of using *CertainTrust* without augmentation with the indicators was attained.

The process of selecting and modelling these three indicators is aimed at illustrating the possibility and basic efficacy of integrating additional knowledge into trust models such as *CetrainTrust*. However, the presented indicators were chosen and integrated into the general framework of *CertainTrust* in a largely heuristic manner. Additionally, they are dependent on the scenario and do not add associativity of arbitrary features and trustworthiness estimates to the estimation process. While certification and insurance are reasonably generic constructs already that can be integrated readily by the trustee, the coalition-based indicator requires active participation by the trustees, e.g., service providers in e-markets. Whether or not an active effort

can be reasonably expected from the trustees to follow a formal coalition-forming protocol is unclear. Furthermore, hardcoding scenario-specific indicators implies considerable modelling effort. This can be alleviated by providing the constructs as part of a modular toolbox of combinable trust tools, in which special extensions, such as those presented in this section, can be bundled.

The following Section 5.2 will explore the use of supervised non-parametric, model-free learning in trustworthiness estimation – thereby, relieving the need for explicitly model and adding feature-associative qualities to learning the trustworthiness of potential trustees. The goal of introducing feature-associativity is to imbue trustworthiness assessment with a higher degree of generalisability, that is, the ability of deriving a trustworthiness estimate by identifying and considering a set of observable features that are typically exhibited by either trustworthy or untrustworthy trustees.

5.2 SUPERVISED METHODS FOR TRUSTWORTHINESS ASSESSMENT

Experience-based Bayesian prediction methods, such as the ones presented in Chapter 3, are the mainstay of computational trust models. However, the reinforcement learning, prevalent in their model design, still offers room for improvement. The reliance on a single type of predictor (either direct or reputation-mediated experience), for instance, leads to poor generalisability. While better generalisability can be reached by direct modification of the trust model and the introduction of new assumptions and model parameters – such as hardcoding indicators of trustworthiness (Section 5.1) – the resulting increase in model complexity is undesirable.

CLASSIFIER ENSEMBLES FOR TRUST By leveraging fusion operators, for instance, those provided by *Subjective Logic* [104] or those discussed in Chapter 4.3.3, p. 114, *classifier ensembles* for trust assessment can be assembled that combine experience-based Bayes estimators with other types of estimators. This permits the integration of sophisticated off-the-shelf estimators and offers an opportunity to leverage advancements in data mining and machine learning. In this section, supervised learning methods, such as *Random Forests*, will be used as consistent non-parametric and model-free supervised learning methods for providing feature-based generalisability in trustworthiness estimation management. Such supervised learning methods can be used for deriving opinions on new agents entering a particular market, based on observable features and experiences a trustee has made with similar agents in the past.

SUPERVISED LEARNING FOR TRUST A number of approaches, particularly stereotyping trust models [31, 128], seek to address the generalisability issue by leveraging supervised learning for trustworthiness prediction. These approaches provide monolithic trust models centred around supervised feature-based prediction. Their focus, however, is on model-building and the presented models require a high discriminatory power of the provided feature set.

Additionally, the distributional assumptions that enable supervised learning methods to build a prediction model depend heavily on the process that generates the data. Here, the influences of a reputation system on the selection and data generation process are often not taken into account, leading to unrealistic distributional assumptions when creating simulated datasets for model validation. For instance, particular distributions of ratings can be observed that mimic the shape of the letter '*J*' (see [91]) in the multinomial case or that favour the positive over the negative feedback category in the binomial case. Supervised learners, however, typically perform best when there sufficient information on all categories and no strong imbalance. Related

approaches [31, 128] partially use theoretical distributional models, leveraging assumptions of Gaussianity, for instance, that do not conform well with the distributions that can be observed in reputation datasets.

Consequently, since the quality of the prediction is predicated on the quality of the data that is presented to the prediction model, trust assessment has to be considered not just from a model-based, but also from a data-driven perspective. To this end, the evaluation presented in this section is based on a real-world dataset of hotel features and ratings, which exhibits distributional properties induced by the data generation process through reputation-based selection. To this dataset, containing more than 3000 hotels, with 33 features for each hotel, several off-the-shelf machine learning algorithms are applied, in order to investigate to what extent the features presented on a hotel booking website encode a hotel's trustworthiness.

In the following, the assumptions and preconditions for performing non-parametric and model-free supervised prediction in trustworthiness assessment (Section 5.2.1, p. 184) will be presented. The hotel dataset is explored and different regression machines are tested on this real-world data in Section 5.2.3. In Sections 5.2.3.1 and 5.2.4, we present and discuss the results and propose a mapping of the estimates to the opinion space representation of commonly used belief logics.

In the latter part of this section, the peculiarities of the dataset, the results of applying supervised learning methods, and describe how to integrate them with existing trust models, e.g., reputation-based methods, by providing a mapping to a belief logic representation, will be investigated.

The work in this Section has been published as a paper [85].

5.2.1 *Approach and Methods*

As opposed to the stereotyping trust models introduced in the related literature [31, 128], the work in this is section is not attempting to present a complete trust model based around a specific supervised prediction method. Rather, the requirements that a supervised prediction approach for trust assessment has to meet will be presented. Additionally, supervised methods will be applied to the dataset and a mapping (in Section 5.2.4) will be provided that enables the integration of the prediction results with existing trust models.

The use of non-parametric, model-free learning methods, following [133], is at the centre of the presented methodology. This is done in order not to be constrained by model assumptions and to ease the burden of excessive parameterising for the user.

The prediction methods that are considered operate in *batch* mode. The data evaluated in Section 5.2.3 are stable with regard to concept

drift – that is, the value of the regressand does not change rapidly. In the given scenario (Hotel Ratings), dataset updates, in the form of newly added hotels and ratings, are comparatively infrequent. Therefore, *online* training methods are not considered here and are relegated to future work. Model update is achieved by retraining the supervised estimators with the entire, updated dataset. Model update is therefore fundamentally equivalent to estimator training, and will not be specifically discussed in further detail.

5.2.1.1 *Pre- and Postconditions*

As a *training precondition*, trust computation based on supervised learning requires a training dataset consisting of $n \in \mathbb{N}, n \gg 0$ records in the form $(\mathbf{x}, y) = (x_1, x_2, \ldots, x_m, y)$. y is the dependent variable, in the case of trustworthiness assessment ideally the true trustworthiness score of a particular trustee, and the vector $\mathbf{x}$ consists of a number $m \in \mathbb{N}^+$ of observable attributes (or *features*) $x_1, x_2, \ldots, x_m$ that are used as input variables. A model-free supervised learning mechanism creates its own prediction model from the data. A trained supervised learning mechanism that is capable of feature-based regression is in the following referred to as a *regression machine*.

As an *assessment precondition*, trust computation requires, once a trained regression machine is available, a feature vector $(x_1, x_2, \ldots, x_m)$ for computing an estimated trustworthiness score $\hat{y}$.

5.2.2 *Consistent Trustworthiness Estimation*

Within the scope of a formal trust model defining trust as a probability, the *postcondition* of the trust computation is, at the least, a probability estimate. The further specifics of this postcondition is determined by the representational model used, for instance for decision making. Thus, when using the *CertainTrust* [174] representational model, we require a probability estimate, as well as a goodness-of-fit (*gof*) measure for determining the certainty parameter.

When estimating probabilities that are to be used in rigorous reasoning, the *consistency* [126] of the estimate is an important prerequisite (see Section 5.2.2, p. 185). A definition of the consistency of estimators is given in Definition 41, p. 186. Consistency of the estimator is not only an important postcondition for probability machines; it also enables us to use an experience-based Bayesian trustworthiness estimate as an estimate for the unobservable trustworthiness of a trustee, i.e., y, as this estimate is consistent itself. Thus, a regression model with a Bayesian trustworthiness estimate for the regressand can be used that will maintain the consistency of the supervised estimator.

Specifically, two distinct cases will be investigated in the following. The first is a regression model in which a trustworthiness score of a particular trustee is available in the training dataset as a probability

score $0 \leqslant y \leqslant 1$. Since this is unobservable, a substitute in the form of a reputation score will be used. In order to meet the consistency requirement for reasoning, this estimate itself should be consistent.

The second case that will be considered is one where only a class label in $\{0; 1\}$ is available in the training data to classify a particular trustee. However, our goal is still to determine an actual probability score $p \in [0; 1]$ for each trustee. For this, we will use so-called *probability machines* [133]; that is, supervised estimators that are known to provide consistent probability estimates from binary regressands.

In the broadest sense, we consider the decision whether or not to trust as a binary classification problem – a truster classifies a trustee as either trustworthy or untrustworthy. In this sense, trustworthiness classification is a discriminatory problem suitably assigned to statistical learning methods. However, in order to satisfy the definition of trust as a subjective probability [64], assigning a class label is insufficient. Rather, the goal in trust assessment is estimating the *probability of class membership*, establishing just *how* likely a particular trustee is to be trustworthy.

Thus, the aim of trustworthiness prediction is to reliably estimate the probability of the trustee acting in a trustworthy manner in the next interaction with the truster, based upon representative input data. Thus, if $y \in \{0; 1\}$ is the outcome of such a future interaction, the goal is to compute a *conditional* probability $P(y = 1|\mathbf{x})$ given the features $\mathbf{x}$. For binary outputs, it follows that $P(y = 1|\mathbf{x}) = E(y|\mathbf{x})$. Both trustworthiness assessment by experience-based Bayesian prediction methods and probability machines leverage this equality in the estimation process.

Obviously, the estimators used for this estimation process should get more precise the more information they are given. That is, they should be *consistent*. Informally speaking, an estimator is consistent, if the error of the prediction converges to zero in the limit *with high probability*.

Formally, the consistency of an estimator can be defined thusly [126]:

Definition 41 (Consistency [126])**.** Let sample $X = (X_1, \ldots, X_n)$ be a member of a sequence corresponding to $n = n_0, n_0 + 1, \ldots$; let $X_1, \ldots, X_n$ be iid according to distribution $P_\theta, \theta \in \Omega$ and $g(\theta)$ be the estimand (the value to be estimated). $\delta_n = \delta_n(X_1, \ldots, X_n)$ is a sequence of estimators.

1. A sequence of random variables X_n defined over sample spaces $(\mathcal{X}_n, \mathcal{B}_n)$ *tends in probability* to a constant c ($X_n \xrightarrow{P} c$) if for every $a > 0$ it holds that $P[|X_n - c| \geqslant a] \to 0$ as $n \to \infty$.

2. A sequence of estimators δ_n of some parameter $g(\theta)$ is *consistent* if for every $\theta \in \Omega$ it holds that $\delta_n \xrightarrow{P_\theta} g(\theta)$.

5.2.2.1 *Experience-based Bayesian Trustworthiness Prediction Model*

To briefly recapitulate[2], state-of-the-art trust models [113] rely on Bayesian prediction models that take experience from past interactions as inputs to compute a probability score. This probability score can be interpreted as the probability that the trustee will act as expected in a future interaction. For the binomial case of trust assessment, we face a classification task with binary class labels for the input (and output) data, i.e., class labels trustworthy and untrustworthy. The data is distributed according to a binomial distribution, generated by repeated Bernoulli trials. In particular, the desired probability value is a point estimate of the Bernoulli distribution's single parameter. This can easily be obtained by computing the expectation value of the binomial distribution's conjugate prior, a beta distribution. For the multinomial case, the class labels and outcome probabilities are distributed according to a categorical distribution and the conjugate prior is a Dirichlet distribution.

Bayesian trust estimators (e.g., [174]) use experience from prior interactions as input. Their output (in the case of binary input variables) is the probability that the *next* interaction with a specific trustee will be a positive one. For the more general multinomial case, the probability estimate accordingly represents the chance that the next interaction will fall into of a specific category, with there being $m > 2$ categories. A fundamentally important quality of Bayesian estimation is its consistency [126].

The basic prediction model of the estimators used in Chapter 3 (see also [108, 174]) is a point estimate of the expectation value of the posterior Beta distribution. That is, if r and s are the sum of positive and negative prior interactions between truster and trustee, the probability estimate – with a uniform prior – is $\frac{r+1}{r+s+2}$. Here, the use of the expectation value of the posterior as an appropriate estimator is due to the equality $P(y = 1|\mathbf{x}) = E(y|\mathbf{x})$. The consistency of this estimator follows from the consistency of the mean as an estimator.

Consequently, experience-based Bayesian prediction yields accurate trust scores, under the assumptions that prior experience is a reliable predictor for future behaviour and that the available prior experience is sufficient – with regard to both quality[3] and abundance – for obtaining a representative point estimate.

The consistency of the estimation method is an important prerequisite for rigorous reasoning. The property of convergence in the limit enables reliable probability assessment of past performance, which is the primary predictor for trustworthiness in computational trust models. Based on the consistency properties of the mean as an estimator of the expectation value, it is reasonable to assume that Bayes-

2 see, Chapter 3

3 Specifically, that the stationarity assumptions hold or non-stationarity is accounted for.

ian trustworthiness estimators represent an adequate regressand for supervised machine learning approaches.

5.2.2.2 *Regression Machines for Trustworthiness Prediction*

A key argument behind the introduction of experience-based computational trust modelling is the scarcity of traditional cues related to trustworthiness in computer mediated interactions [113]. Such cues are equivalent to indicators of trustworthiness – defined in the preceding Section 5.1 (Definition 37, p. 162) as a feature or set of features that a trustee possesses that are supposedly representative of its trustworthiness. While *traditional* cues learned from interactions in brick-and-mortar environments often cannot be applied to online interactions, modern online services expose a wealth of observable features. These can form the basis for learning new cues, which in turn can provide estimators for computational trust assessment by using supervised learning that perform better, for instance under the absence of direct experience with specific trustees.

Data mining approaches for exploiting (potentially highly dimensional) feature spaces for probability estimation tasks are numerous, as evidenced by a large proliferation of regression mechanisms (see, for instance, [47, 182]). Parametric models, such as logistic regression, are traditionally applied there. However, traditional parametric models suffer from drawbacks that limit their use in trust assessment in computer mediated interactions. In particular, parametric models have to be specifically fitted to the problem they are to address. In order to avoid model misspecification, predictors and supposed interrelations have to be input correctly. Additionally, parametric models make assumptions on the distribution of the data used in the regression, particularly with regard to the Gaussianity of the data. This limits their use considerably considering the scalability and flexibility required in data-rich environments where features can exhibit different scale types, dimensionality, distributions and correlation structures [133].

Model-free, non-parametric regression machines support the robust estimation of conditional probabilities from feature sets of different scale types and potentially high dimensionality. They make no distributional assumptions for the vector of features, make no restrictions on the length of the feature list, and do not rely on a specified model as a starting point [133]. They do, however, require more data than the parametric models for the model building process inherent in non-parametric methods. While in parametric methods, the model is predefined and has to be parameterised, in model-free non-parametric methods, the model is created from the data itself. The focus in this section is on model-free, non-parametric regression machines.

In order to allow for *robust* probability estimation and thereby enable rigorous and meaningful inferences with regard to the trustworthiness of a trustee, consistency of the regression model is important. In the following, an experience-based Bayes estimate will be used as a regressand, When using such an experience-based Bayes estimate of the trustworthiness score as regressand, consistency is inherent in the consistent Bayes estimator for the regressand. For the regression task, therefore, the regressand parameter for the training step is considered to be sufficiently exact.

Additionally, particularly when using a class label, instead of an already consistent estimate of the trustworthiness score, the supervised estimator itself has to be consistent. Malley et al. [133] term consistent non-parametric and model-free probability estimators that estimate the conditional probability function for a binary outcome as *probability machines*. The selection of machine learning methods used will consequently be limited to those for which consistency has been shown in the related work; several different probability machines will be applied to the task of trustworthiness assessment, namely, Random Forests [25], k-Nearest Neighbour [16] approaches and Decision Trees [26, 166]. The regression model and the different probability machines will be briefly introduced in the following.

REGRESSION MODEL Following Malley et al. [133], the probability estimation problem constituted by trustworhiness estimation will be treated as a *non-parametric regression* problem. Thus, the regression machine will serve to estimate the non-parametric regression function $f(\mathbf{x}) = E(y|\mathbf{x}) = P(y = 1|\mathbf{x})$, where $\mathbf{x}$ is a vector of features (regressors). This requires no data input from the user to specify and tune the model. However, non-parametric, model-free methods have the disadvantage of requiring considerably more data for model creation than those for which the model is predefined. The presented model assumes binary feedback categories, which is supported by the real-world hotel data set analysed in this section. The general supervised prediction model and the notion of probability machines as used in [133] can be extended to the multinomial case, for instance by adapting work by Kruppa et al. [122].

In many application scenarios feature sets of predictors that serve as regressors can be obtained by the user with relative ease. Methods of web data extraction, for instance, can be employed for gathering features associated with a potential trustee that can serve as regressors. The hotel data set, presented in Section 5.2.3, p. 192 and used in this chapter, offers a number of potential features. However, the *true* regressand, that is the intrinsic trustworthiness of the trustee, is generally an unobservable variable in real-world applications. In its place, a trustworthiness estimate from an experience-based Bayes estimation method can be used. Ideally, this estimation method is a ro-

bust reputation-based trust model, such as [108, 174]. However, due to the mostly academical nature of these works and their absence in real-world applications (such as e-commerce sites), simpler, more widely-used basic reputation systems will have to be substituted instead. When testing supervised estimators in pure regression mode, this substitution is direct. When using the estimators as probability machines with a binary regressand, a binary dichotomisation of the reputation score serves as the regrassand. Specifically, the estimators are created from a real-world data set, described in 5.2.3, p. 192.

RANDOM FORESTS Random forests [25] are non-parametric ensemble classifiers consisting of a multitude of decision trees. They are generally considered to be fast and and accurate classifiers that offer considerably better performance than single trees [16], for instance, CART [26] or M5 [166].

Random forests have several strengths that make them theoretically well-suited to trustworthiness assessment. In particular, they can handle high dimensional feature spaces of different scale types, with little user input. Thus, they can be presented with arbitrary sets of feature vectors that result from web data extraction, without requiring user-driven feature selection or model specification. Additionally, they typically provide robust estimates, even under conditions of missing data. Conveniently, random forests perform rudimentary error estimation using an out-of-bag (OOB) method[4][25] during the learning process and give estimates of which features are important in the classification or regression tasks.

In classification tasks, the output of a random forest is the mode of the classification outputs of its constituent classification trees. Instead of outputting a class label, the random forest can also return an estimate of the conditional probability $P(y|\mathbf{x})$. As we are concerned with *probability estimation* of (binary) classes, the probability estimate can be obtained by computing the proportion $\frac{|y=1|}{|y=0|+|y=1|}$, averaged over all constituent trees, when running the random forest in classification mode. In regression mode, the random forest consists of regression trees instead of classification trees. Thus, the probability estimates are averaged over the regression results of the individual trees instead. For the prediction of hotel ratings (section 5.2.3), we will use random forest estimators in classification and regression mode, termed *classRF* and *regRF*.

The consistency of random forests has been shown by Biau et al. [17]. [25, 133] provide a detailed description of random forest bootstrapping and classification procedures.

Random forests also exhibit weaknesses. For instance, they do not generate easily interpretable classification rules. Consequently, they

4 Therefore, they do not necessarily require dedicated cross validation to control overfitting.

are not suitable for creating the explicit groups required for rule-based stereotyping approaches to trust assessment. The intuitive human interpretability of the correlation of feature vectors to class estimate is therefore limited.

K-NEAREST NEIGHBOUR k-Nearest Neighbour (*k-NN*) estimators are a special case of kernel density balloon estimators [67]. The (simplified) classification process is intuitive: An unlabelled sample is classified by comparing its feature vector to labeled samples from a training set and choosing the k closest according to an appropriate distance metric. The class of the unlabelled sample is estimated by determining the mode of the k labels of the labeled neighbours. In a regression model with a continuous regressand, the mode can, for instance, be replaced by an inverse distance weighted average function.

Breiman [24] introduced a variation of nearest neighbour classifiers that combines several k-Nearest Neighbour into an ensemble classifier, using *bagging* (bootstrap aggregating). This is analogous to the formation of random forests from decision trees. Thus, the output of the bagged *k-NN* (*b-NN*) is the mode of its constituent *k-NN* estimators for a classification task. A probability estimate can be obtained in the same manner as for the *classRF* random forest [133].

DECISION TREES In recent publications dealing with the application of machine learning to trustworthiness assessment tasks [31, 129], decision trees have been used for classification tasks. There are several decision tree algorithms that can perform regression and are suitable for trustworthiness assessment. Specifically, we will test CART [26] and M5 [166] decision tree algorithms on the dataset.

Decision trees offer white box behaviour and interpretability of the generated models. They are also reasonably robust, performant and can deal with different scale types as input data.

Another popular class of estimators, support vector machines (*SVM*), are omitted because they do not guarantee universal consistency [133].

In Section 5.2.3 supervised methods are tested on a real-world dataset – with regard to their capability to predict reputation scores from features. Intentionally this evaluation of the prediction methods is not done on synthetic data. The power of the machine learning methods described above, i.e., *random forests*, *k-NN* estimators and *decision trees* is well-established. Generating synthetic data to show the discriminatory qualities of these methods would thus be only a replication of work. For an application of probability machines to benchmarking datasets, the interested reader is referred to [133].

Scale Type	Feature
Nominal	ID, City
Binary	*Payment Options*: Master, Visa, AmEx
	Hotel Ammenities: Laundry Service, WiFi, Restaurant, Bar, Bistro and Cafe, Steam Bath, Elevator, Special Access, Gym, Sauna, Solarium
	Room Ammenities: Telephone, TV, Radio, AC, Safe, Minibar, Desk, Hair Dryer, Bath Tub
Ordinal	Hotel Stars
Ratio	Aggregate Recommendation, Number of Recommendations
	Distances to next: Airport, Highway Access, Railway Station, Commuter Station
	Number of Rooms: Total, Single, Double
	Price

Table 10: Scale Types and Features for the Hotel Dataset

5.2.3 *Application of Supervised Predictors to Data*

Hotel booking and ranking sites represent a real-world application of reputation systems that combine both electronic availability of the reputation data, as well as physical service provisioning in a mature and regulated market. The records furnished by hotel booking sites actually guide real customers to make a trust decision and, through their rating feature, provide a feedback mechanism. They provide the user not only with reputation scores for hotels, but also with collections of features, that are standardised, complete and verifiable to some extent. The physical nature of the service provisioning and the correspondingly required monetary collateral (e.g., costs of realty, furnishings, personnel, etc.) justify assumptions of slow concept drift and market persistence of individual hotels.

In order to test regression machines for trustworthiness assessment, a dataset of 3,006 hotel records for hotels in 9 major European cities from a German hotel booking site was acquired. Each record consists of an ID, an aggregated rating score, the number of individual binary ratings that were aggregated into the rating score, as well as 33 features of various scale types (Table 10).

The aggregate rating score represents the average probability of a good outcome in a binary setting; that is, when rating a hotel, raters were asked '*Would you recommend this hotel?*' and could answer either *yes* or *no* – individual ratings, therefore, are binary. Rating aggregation into an aggregate recommendation score is achieved via simple averaging. Ratings are only available as aggregate scores, that is, no time series of individual ratings was available. Furthermore, raters were only able to rate hotels that they had booked through the booking site. This curbs malicious or fake positive reviews by increasing the transactional cost a false review carries [139, 140], supposedly making the reviews more honest overall and leading to a more exact regressand.

Overall, raters contributed 199,168 ratings, of which 151,868 ($\approx$ 76%) were positive and 47,300 ($\approx$ 24%) were negative ratings. Of the 3,006 hotels in the dataset, 356 ($\approx$ 11.8%) have not been rated yet. Of those 2,650 hotels that have been rated, the mean number of ratings per hotel is 75.16 – the median, however, is considerably lower at 25 (for a summary, see table 11). Figures 28a, 28b show histogram information of aggregate recommendations, clearly displaying the peakedness of the empirical distribution and the effect of the excess positive individual and aggregate ratings (see also Table 11). The histograms show the count for hotels in twenty 5 percentile bins, that is, in categories 0 to 5 per cent positive ratings, 5 to 10 percent positive ratings, and so on. Additionally, unrated hotels are accumulated on the left hand side of the histograms. The particular choice of bin width was made to provide a visualisation that is both sufficiently clear and detailed. Diagrams 28a to 28d are reproduced, in a larger format, in Appendix E.

Figure 28c shows a long-tailed distribution of the number of recommendations per hotel, i.e., a small number of hotels have a high number of recommendations, while the vast majority of hotels have a comparatively small number of recommendations. Figure 28d plots the distribution of the recommendation score against the number of recommendations. The distribution evident in these figures hints at *preferential attachment* processes that are induced by the decision making and feedback mechanisms of the reputation system.

In next Section (5.2.3.1), the off-the-shelf regression machines described in Section 5.2.1 are applied to the hotel dataset. The general learning procedure leverages the non-parametric regression function $f(x) = E(y|\mathbf{x}) = P(y = 1|\mathbf{x})$, where $\mathbf{x}$ is a vector of features (regressors). The aggregate recommendation score is used as regressand, while the 33 features listed in Table 10 (omitting *ID* and *Number of Recommendations*) will serve as regressors. The aggregate recommendation score is assumed to be an adequate surrogate for the unobservable true trustworthiness of each trustee (i.e., hotel), which is justified by the arithmetic mean being a consistent and stable estimator.

	Min.	**1st Qu.**	**Median**	**Mean**	**3rd Qu.**	**Max.**	**SD**	**Skew**	**Kurt**
Number	2	9	25	75.16	78	1531	132.40	4.03	23.08
Score	0.0	0.65	0.75	0.73	0.83	1.0	0.138	-0.89	1.38

Table 11: Distribution of number of recommendations and recommendation score.

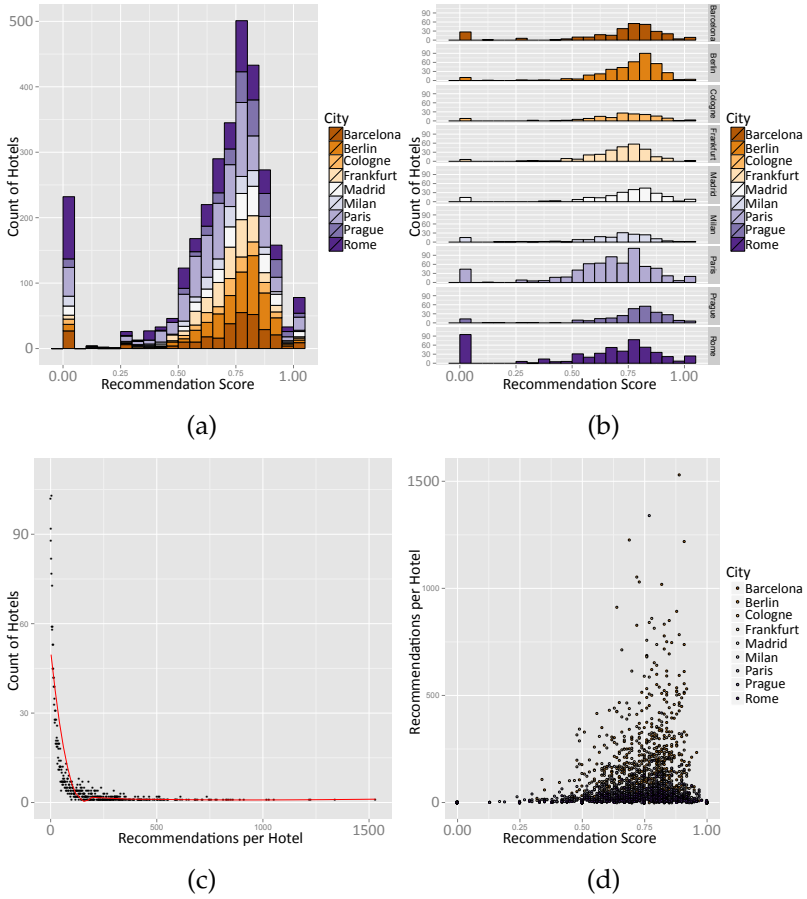


Figure 28: Aggregate recommendations in the hotel dataset.

5.2.3.1 *Results*

The estimators that were introduced in Section 5.2.1 are designed to reveal correlations between the features and the regressand variable. In order to see whether or not such correlations exist and can be detected by the supervised estimators, the estimators are tested in two different settings. First (Section 5.2.3.2), the random forest, *k-NN*, CART and M5 decision tree algorithms are applied in a regression scenario with the aggregate recommendation scores as unmodified regressands $y \in [0; 1]$. In addition, logistic regression [89], a very basic supervised learning method, was also applied to the data to provide a baseline. Second (Section 5.2.3.3), the probability estimation capabilities of the estimators were evaluated in a classification scenario (i.e., in a dichotomous regression scenario with values 0 or 1, with the estimators operating as probability machines). For this, we generated dichotomous outcomes from the aggregate recommendation scores. For each hotel, a new dichotomous response variable y was computed by using a binomial random number generator with the hotel's recommendation score as the corresponding probability. Random forests, *k-NN*, *b-NN*, CART and M5 decision tree estimators were

trained using the new binary response variable and the 33 features of the hotel dataset as regressors. The estimators were *not* presented with the recommendation scores or the number of ratings per hotel.

Estimator training was repeated 50 times, the reported results for the quality of the estimators represent the mean of the repeated training process. During each repeat, the dichotomisation procedure was reapplied.

In both cases, the area under the curve (*AUC*) was computed against the dichotomised binary response, based on the receiver operating characteristics (*ROC*) [53]. The *ROC* plots the fraction of true positives out of the total actual positives, i.e, the true positive rate (also called *sensitivity* or *recall*), against the fraction of false positives out of the total actual negatives, i.e., the false positive rate (which is $1-$*specificity*), at various cutoff levels. Thus, the *ROC* is a measure of the performance of an estimator in a binary setting. The integral underneath the *ROC*, i.e., the *AUC*, is a single value representation of the estimator performance.

For goodness-of-fit (*gof*) evaluation, random bootstrap samples were drawn, and used to train the estimators on the in-bag samples and evaluate the performance using out-of-bag (*OOB*) samples. Additionally, 10-fold cross validation (*CV*) was performed to check for overfitting. None of the estimators exhibited tendencies towards overfitting the data and the goodness-of-fit *gof* did not vary noticeably between random forest *OOB* estimates, standard holdout and *CV*. We evaluated *gof* according to several standard error measures (see Table 12) [150] based on the difference between the estimates $\hat{P}(y = 1)$ and the recommendation score, which we assume to represent the true trustworthiness $P(y = 1)$. The measures presented in Table 12 are (according to Moriasi et al. [150]):

- *Mean Error* (ME): the mean of the differences between the observed values, i.e., recommendation scores, and their corresponding estimates;
- *Mean Absolute Error* (MAE): the mean of the absolute difference between the observed values, i.e., recommendation scores, and their corresponding estimates;
- *Mean Squared Error* (MSE): the mean of the squared difference between the observed values, i.e., recommendation scores, and their corresponding estimates;
- *Root Mean Squared Error* (RMSE): the square root of the mean of the squared difference between the observed values, i.e., recommendation scores, and their corresponding estimates – the RMSE, conveniently, has the same units as the estimand;

- *Normalised Root Mean Squared Error* (NRMSE): the RMSE normalised by dividing by the standard deviation of the observations – in Table 12 given in percent, i.e., multiplied by 100.

Aside from these measures, describing the first and second moments of the error function, the following standard measures are also given:

- *Percent Bias* (PBIAS): a measure giving the difference between the estimators expected value and of the observations, multiplied by 100 as percent; it gives the tendency of the estimates to be either larger or smaller than the observations;
- *Ratio of RMSE to the Standard Deviation of the Observations* (RSR): the RMSE normalised by the standard deviation of the observations, providing an error statistic ranging from the optimal value 0 to large values. A value close to 0 signifies a good model fit;
- *Ratio of Standard Deviations of Estimates and Observations* (rSD): an error index indicating the dispersion of the estimates relative to that of the observations;
- *Nash-Sutcliffe Efficience* (NSE) and *Modified NSE* (mNSE): normalised statistics that indicate the ratio of noise to information;
- *Modified Index of Agreement* (md): a standardised measure that indicates the degree of agreement between estimates and observations, ranging from 0 (no agreement) to 1 (complete agreement).

Mathematical definitions of the *gof* measures is given in Appendix F.

In the following, the model fit will mainly be discussed in terms of the *MSE* and its derivatives. The *mean squared error* (*MSE*), that is, the average squared difference between estimates and observations, provides an analytically tractable foundation for measuring the goodness of an estimator. It is of particular interest, because it simultaneously measures the variability of an estimator, i.e., its *precision*, as well as its bias, i.e., its *accuracy* [35].

In the following let $O = (o_1, o_2, \ldots, o_n)$ be a vector of observed values and $\hat{S} = (\hat{s}_1, \hat{s}_2, \ldots, \hat{s}_n)$ a vector of corresponding estimates. Then the *MSE* is defined as follows:

Definition 42 (Mean Squared Error (MSE))**.**

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(\hat{s}_i - o_i)^2$$

However, as in Chapter 3.1.2, p. 52, we would like to represent the quality of the estimator with a parameter that is expressed in the same units as the estimand value. Using the *MSE* as a basis, the *root mean squared error* (*RMSE*) provides a goodness-of-fit measure that meets this requirement.

Definition 43 (Root Mean Squared Error (RMSE))**.**

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{s}_i - o_i)^2}$$

Furthermore, the distribution of the estimand parameter within the data should be accounted for. If the deviation of the estimand parameter is low, that is, if the realisations of estimand parameter are considerably clustered, the generalisation qualities of the estimator cannot be guaranteed. This problem is evident in the hotel data set, Figure 28, p. 195, where the observations are strongly clustered around a value of 0.75.

Because the supervised estimators used for the estimation task typically optimise the *MSE*, a strongly unbalanced data set – that is, a data set in which one category is considerably over-represented– such as the hotel data set (considerably more positive 1-ratings than negative 0-ratings), can lead to good *MSE* scores but poor generalisability. This is caused by the estimators learning the distribution of the estimand parameter instead of the discriminatory capabilities of the presented feature set. Particularly when paired with feature sets with low discriminatory power, the unbalanced nature of the data and the resulting low variance of the estimand parameter make it advantageous to the estimator to place all estimates close the mean of the observations. In order to account for this behaviour, the *RMSE* can be normalised by a measure for the variability of the estimand, for instance the standard deviation of the observations. This yields the *normalised root mean square error (NRMSE)*.

Definition 44 (Normalised Root Mean Squared Error (NRMSE))**.**
Let $\text{sd}(o_i)$ denote the standard deviation of the observed values O.

$$\text{NRMSE} = \frac{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{s}_i - o_i)^2}}{\text{sd}(o_i)}$$

An alternative formulation of the NRMSE uses a different normalisation criterion: Let o_{max} be the largest, o_{min} the smallest element of O.

$$\text{NRMSE} = \frac{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{s}_i - o_i)^2}}{o_{max} - o_{min}}$$

Because the standard deviation (sd) is a more robust measure of the variability of the estimand, it is preferred as a normalisation factor.

The *NRMSE* is a more conservative a measure than the *RMSE*. Tables 12, p. 200, and 14, p. 202, illustrate this for the hotel data set. For this particular data set, when comparing the *RMSE* to the *NRMSE*, given as the percentage $\text{NRMSE}\% = 100 \cdot \text{NRMSE}$, the difference is considerable, with the two error estimates varying by a factor of ≈ 7. Combined with the shape of the absolute error curve in Figure 29b, p. 203, this leads to the conclusion that the available data and the discriminatory power of the feature set are insufficient for making reliable standalone trustworthiness estimates. For the integration into the prior of an experience-based Bayesian Trust model, the following mapping is provided; in principle any measure of the prediction error can be used. For the reasons outlined above, the *NRMSE* was chosen in Defintion 46.

Random forest estimators were applied in regression mode (*regRF*, to both recommendation score and class label regressands) and classification mode (*classRF*, to class label regressand). For each of these, two distinct configurations were chosen: one that guarantees consistency (according to [133]), in which individual trees were not fully grown, and one that grows the individual trees to their full extent, according to the default settings [25] for *regRF* and *classRF*. Note that in the latter case, universal consistency of the random forest estimator cannot be guaranteed [133].

5.2.3.2 *Regression to Recommendation Score*

When running the estimators in regression mode, the regressand is a recommendation score in $[0; 1]$. The estimate is also in $[0; 1]$. Goodness of fit measures were computed accordingly from these two.

The results of applying the regression machines can be seen in Table 12, in terms of various goodness of fit measures (for a documentation of the measures, see [208] and Appendix F). The normalised root mean square error (NRMSE, see definition 44) indicates that the random forest estimators perform marginally better than the decision trees. As per the mapping presented in the discussion section (Section 5.2.4), a prediction is considered informative if the *percentage* NRMSE (NRMSE%) is smaller than 100. While all tree-based estimators (*regRF*, *M5*, *CART*) achieve an $\text{NRMSE}\% < 100$, nearest neighbour and logistic regression return no informative results.

When considering the *AUC*, as per Table 13, the random forests outperform the other estimators. However, the margin between the different methods is small, and the overall performance of all methods is slightly but statistically significantly better than random guessing, which would correspond to an *AUC* of 0.5).

	ME	MAE	MSE	RMSE	NRMSE %	PBIAS %	RSR	rSD	NSE	mNSE	d	md
regRF (consistent)	0	0.1	0.02	0.13	**91.8**	0	0.92	0.29	0.16	0.11	0.43	0.32
regRF (default)	0	0.09	0.02	0.12	**89.6**	-0.6	0.9	0.44	0.2	0.14	0.56	0.41
M5	0	0.09	0.02	0.13	**91.8**	0	0.92	0.44	0.16	0.11	0.53	0.38
CART	0	0.1	0.02	0.13	**94**	0	0.94	0.39	0.12	0.08	0.47	0.35
k-NN	0	0.11	0.02	0.14	**103.4**	-0.3	1.03	0.52	-0.07	-0.02	0.45	0.34
logit	0.29	0.33	0.17	0.42	**301.2**	39.3	3.01	2.37	-8.07	-2.12	0.39	0.26

Table 12: Average goodness-of-fit of regression to recommendation score (for a documentation of the measures, see [208]).

	avg AUC	**MIN**	**MAX**	**± SD**
regRF (cons)	0.590***	0.563	0.604	± 0.012
regRF (def)	0.585***	0.565	0.599	± 0.014
M5	0.582***	0.565	0.6	± 0.012
CART	0.56***	0.543	0.575	± 0.01
k-NN	0.547***	0.543	0.55	± 0.01
logit	0.582***	0.563	0.603	± 0.014

Table 13: Average classification performance with recommendation score as regressand (***: p value (95 % confidence interval) of one-sided Wilcoxon test, AUC prediction vs. guessing, i.e. $\mu = 0.5$, $p < 0.001$).

5.2.3.3 *Regression to Class Label*

In the second scenario, the regressands used for training the estimators are class labels in $\{0; 1\}$, however the resulting estimates are still estimates in the interval $[0; 1]$. In order to produce class label regressands, the recommendation scores of the hotel data set were dichotomised. The procedure for doing so is straightforward: as the recommendation score is known, as well as the total number of recommendations for each hotel, a corresponding number of $\{0; 1\}$ ratings can be generated. That is, per hotel there exist as many dichotomised ratings as the total number of recommendations for that hotel. Multiplying the recommendation score by the total number yields dichotomised 1-ratings, while subtracting the number of 1-ratings from the total number of recommendations yields the dichotomised 0-ratings. Learning from class labels is useful in cases in which no distinguishing identifiers are available, aside from the feature vectors.

When operating the estimators as probability machines, results of the probability estimation (Tables 14 and 15) are qualitatively broadly similar to those of the regression machines in Section 5.2.3.2. Goodness-of-fit of the probability estimates and classification performance (as *AUC*) are even weaker, however. Only the consistent *regRF* and the two nearest neighbour approaches achieve a $\mathrm{NRMSE}\% < 100$.

Figure 29 shows the predictive performance and absolute error of the best performing (in terms of *AUC*) estimator, a consistent *regRF* trained on recommendation score regressands. The distribution of the prediction versus the actual recommendation score and the distribution of the error indicate the limited ability of the estimator to create a good prediction model. Predictions are centred around the mean recommendation score, thereby decreasing the goodness of the prediction the further the actual recommendation score deviates from this mean. Majority class undersampling was performed to check if this was solely induced by the distribution of the recommendation score. However, undersampling did not lead to improved perfor-

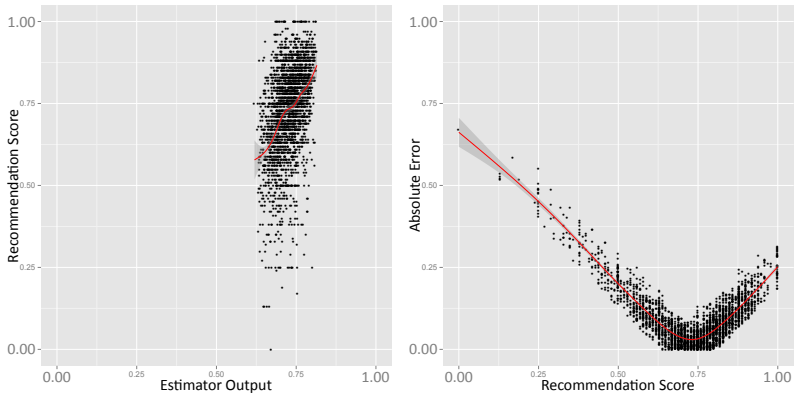
	ME	MAE	MSE	RMSE	NRMSE %	PBIAS %	RSR	rSD	NSE	mNSE	d	md
regRF (consistent)	0	0.1	0.02	0.13	**94.2**	-0.1	0.94	0.36	0.11	0.08	0.45	0.33
regRF (default)	-0.02	0.12	0.02	0.15	**110.6**	-2.9	1.11	0.75	-0.22	-0.11	0.52	0.38
classRF (consistent)	0.26	0.26	0.09	0.29	**212.8**	35.6	2.13	0.13	-3.53	-1.43	0.39	0.29
classRF (default)	-0.01	0.11	0.02	0.15	**107.6**	-1.7	1.08	0.7	-0.16	-0.07	0.52	0.38
M5	0	0.1	0.02	0.14	**102.2**	-0.1	1.02	0.59	-0.04	0.03	0.49	0.37
CART	-0.46	0.46	0.23	0.48	**347.6**	-62.8	3.48	0.17	-11.08	-3.3	0.3	0.19
k-NN	-0.01	0.1	0.02	0.13	**96.8**	-1.2	0.97	0.28	0.06	0.04	0.36	0.27
b-NN	-0.01	0.1	0.02	0.13	**96.8**	-1.1	0.97	0.3	0.06	0.04	0.37	0.28
logit	0.3	0.41	0.35	0.59	**427.8**	41.5	4.28	3.8	-17.3	-2.84	0.27	0.23

Table 14: Average goodness-of-fit for regression to class label (for a documentation of the measures, see [208]).

	avg AUC	MIN	MAX	± SD
regRF (cons)	0.568***	0.552	0.585	± 0.013
regRF (def)	0.547***	0.523	0.579	± 0.019
classRF (cons)	0.529***	0.503	0.545	± 0.012
classRF (def)	0.55***	0.527	0.579	± 0.02
M5	0.554***	0.523	0.584	± 0.019
CART	0.529***	0.505	0.544	± 0.012
k-NN	0.548***	0.529	0.564	± 0.014
b-NN	0.541***	0.505	0.564	± 0.024
logit	0.557***	0.535	0.583	± 0.016

Table 15: Average classification performance with class label as regressand (***: p value (95 % confidence interval) of one-sided Wilcoxon test, AUC prediction vs. guessing, i.e. $\mu = 0.5$, $p < 0.001$).

mance, leading to the conclusion that the features presented to the probability machines have too little discriminatory power.



(a) Prediction (b) Absolute Error

Figure 29: Predictive performance and absolute errors for regression random forest (regRF, consistent, ntree=10%).

5.2.4 *Discussion of Prediction Results*

The dataset presented in Section 5.2.3 illustrates peculiarities that are caused by the presence of reputation systems in service selection. The data exhibits a strong disproportion of positive ratings over negative ones (Figures 28a and 28b). Assuming that ratings are, for the most part, authentic, this may be attributed to two main reasons.

1. The type of service provided is physical in nature, rather than virtual, and has a long and established tradition, and is well-regulated by social norms, as well as economic and legal bodies.

Thus, providing a service as advertised is strongly encouraged by the environment of service provisioning. At the same time, there are established expectations what a customer can expect from the service provider/hotelier, leading to positive expectation confirmation. Simply put, providing a physical service as advertised is simply the social and legal norm, while at the same time the customer knows what to expect from a 3-star hotel at a given price point.

2. More interesting, from a data-centric perspective, is a tendency towards preferential attachment that is visible from the data. Considering Figures 28c and 28d, we can observe
   a) a long-tailed distribution signifying that only a small number of hotels have many ratings, reminiscent of a power law distribution; and
   b) high numbers of ratings are considerably more frequent among hotels with higher recommendations scores.

   Because hotels with good ratings are preferentially selected – as a risk minimisation strategy – and because hotels with a good rating can be considered to be more likely to provide satisfactory service, reputation systems contribute to the skewed distribution observable from the data. By design, reputation systems dissolve the independence of service selection and feedback – a fact that is both indicative of and contributes heavily to their success as a soft security control instrument. Well-behaving providers are rewarded by building a good reputation and attracting more customers, while badly performing providers are effectively eliminated from the selection process.

Thus, the dataset reflects the success of a functioning reputation system in a real-world application scenario, in which transaction costs are non-negligible. At the same time, however, the effects of preferential attachment that are driven by the reputation system also pose challenges. Exploitative service selection is encouraged over explorative selection, which leads to established markets and market entry issues for new hotels. Not only that, but because *presumably* bad providers are very quickly eliminated from the market, by not being selected, *and* because these presumably bad providers only have a low number of bad ratings, feature-based trustworthiness prediction methods are limited in their effectiveness. The number of negative samples is simply not sufficient to build accurate predictive models.

Consequently, the features presented on the selected hotel booking website encode a hotel's trustworthiness to a very limited degree. This limits the usefulness of stereotyping approaches in service selection scenarios, because the data foundation that is used for machine learning is necessarily skewed by the selection process. However, the performance of the regression machines is still significantly (Tables

13, 15) better, statistically speaking, than pure guessing and therefore can (and should) be harnessed.

5.2.5 *Supervised Estimation to Opinion Mapping*

The goodness-of-fit of the supervised estimators evaluated in Section 5.2.3.1 does *not* warrant building a standalone trust management system around them. The features of the hotel dataset do not provide sufficient discriminatory power to build accurate models from the skewed data and do not yield reliable trust scores. However, the regression and probability machines still provide if not an accurate fit of the trustworthiness, then at least an *indication* of how trustworthy a particular hotel is. As such, they can still be of value within a trustworthiness estimation *ensemble* (Definition 45). They can be used in a supplementary role, for instance as input to the base rate or initial expectation of an experience-based Bayesian model.

Definition 45 (Trustworthiness Estimation Ensemble)**.** A *trustworthiness estimation ensemble* is any estimation mechanism that combines several individual estimators to produce a trustworthiness estimate.

The meaningful combination of different trustworthiness estimators and the logical inference over their output require a framework for reasoning. *Subjective Logic* [113] is a popular choice for reasoning under uncertainty that is inherent in the estimation process. *CertainLogic* [175] is more recent but similar framework, which is derived from and fully isomorphic to *Subjective Logic*.

We model the integration of trust estimating regression machines with other estimators, e.g., reputation-based trust models, using *CertainLogic*. This choice is governed primarily by the fact that the opinion representation of *CertainTrust*, which *CertainLogic* extends, corresponds more intuitively to the outputs and error estimates of the regression machines. Choosing *CertainLogic* over *Subjective Logic* should not be understood as a reflection on the capabilities of each; rather, the use of the *CertainTrust* opinion representation is hoped to ease understanding.

In Chapter 4.3, extensions of those *CertainLogic* operations that are required for enabling trustworthiness estimation ensembles have been presented. More complex ensembles that necessitate the use of logical operators, such as *AND* and *OR*, can be assembled by leveraging the bijection between the *CertainTrust* opinion representation and the evidence space, represented as sufficient statistics over the samples, e.g., *sum of successes* and *sum of failures* for binary samples.

CertainLogic is derived from *Subjective Logic*, which is rooted in belief theory [183]. As such, it allows not only for the modelling, combination and inference over probabilities, but over *CertainTrust opinions*. Recall that *CertainTrust* opinions allow explicitly expressing any pos-

sible *uncertainty* regarding the probabilities (see Chapter 3). Binomial *CertainTrust* opinions are ordered triples $\omega = (t, c, f)$, where:

- $t \in [0; 1]$ is a *probability estimate* that $y = 1$.
- $c \in [0; 1]$ is a *certainty estimate* that the probability estimate t is correct.
- $f \in [0; 1]$ is a *dispositional parameter*, modelling an a-priori assumption, thus encoding a Bayesian Prior.

In experience-based Bayesian trustworthiness prediction, such as *CertainTrust* [108, 174], the probability estimate t generally corresponds to a point estimate of the expectation value of the posterior Beta distribution, such as the proportion of good ratings ($y = 1$) to all ratings a truster has with regard to a specific trustee (see Chapter 3). The certainty estimate c is typically a function of the number of such ratings. In Chapter 3.1.2, a more sophisticated pair of certainty estimators (Definitions 10 and 12), was introduced that also takes into account the variance of the posterior Beta distribution. Establishing the certainty estimate in this manner is made possible by leveraging basic model assumptions of Bayesian prediction, in particular the convergence of the posterior mean to the true expectation value with increasing evidence.

Experience-based Bayesian trustworthiness assessment represents a fully realised model for the prediction of future performance based on past performance. This model is a sound statistical prediction model, for which estimators for expectation value and variance are readily available. The convergence of the estimate to the true value, and therefore the certainty parameter, can be derived from the variance of the posterior distribution. Due to the rigid nature of the employed statistical model, experience-based Bayesian trustworthiness assessment does not lend itself to feature-based generalisation, however.

Conversely, regression machines try to create a prediction model from the data they is presented during training. This prediction model is then used to make predictions based on a presented feature vector, thereby providing generalisation. How well such a model generalises is highly dependent on the data that is available for training. Consequently, the certainty parameter c cannot be estimated by leveraging convergence properties that can be derived from model assumptions.

When using regression machines, mapping the probability estimate t is trivially achieved by using the prediction value, as in Bayesian models. Certainty estimation, however, has to be done in a different manner, due to different characteristics and purposes of the prediction paradigm.

5.2.5.1 *Estimator Quality as Certainty*

Since the main goal of introducing supervised learning methods is to add feature-based generalisation to trustworthiness assessment, the quality of the generalisation has to form the basis for the certainty estimation. A number of error measures are conventionally used to express the performance of supervised learning approaches. As a basis, any increasing function of the absolute distance between observations and corresponding predictions serves as a measure of the goodness of an estimators [35].

In binomial trustworthiness assessment, which essentially represents a probability estimation task, the scale of estimand is identically distributed for any individual trustworthiness estimation task, that is, the scale of the trust score does not differ when estimating the trustworthiness of trustee P_1 from that estimating the trustworthiness of another trustee P_2; neither does it change from truster A to truster B. Therefore, the use of a scale-dependent quality measure is possible, as the estimand is always on the scale $[0; 1]$. In other words, comparisons are always made with regard to the same variable.

Definition 46 (Supervised Estimator to *CertainTrust* Opinion Mapping)**.** The mapping from estimator output to *CertainTrust* opinion space is given as:

- $t = P(y = 1|\mathbf{x})$
- $c = 1 - (\min(NRMSE, 1))$
- $f \in [0; 1]$, representing prior information.

The mapping of $c = 1 - (\min(NRMSE, 1))$ may be considered somewhat ad-hoc. In principle any measure of the *gof* that scales between 0 and 1 may be used. However, since the *NRMSE* is a reliable, widely used and standardised measure of the goodness of a model fit, it is a reasonable choice. A closer investigation of which particular measure to choose in order to guarantee optimal performance of the trustworthiness prediction is relegated to future work.

The provided mapping enables the integration of regression machines in trustworthiness assessment ensembles. Using different fusion operators (see [77, 104]), different estimation paradigms can be flexibly combined, thereby enabling ensembles that can leverage the respective strengths of the different estimation paradigms.

5.2.5.2 *Supervised Prediction as Initial Trust Value*

As pointed out earlier, the discriminatory power of the features exhibited in the hotel data set is limited. This makes the use of stereotype-based trustworthiness estimation – as the only way to compute the trustworthiness t of a trustee – for this data set infeasible. However, the information contained within the data set can still be leveraged by

using the feature-based prediction as input for the initial trust value f of the *CertainTrust* model.

In order to determine the effect of instantiating the *CertainTrust* expectation value computation, $E = c \cdot t + (1 - c) \cdot f$, using the feature-based prediction as input for the parameter f, an experiment was set up as follows. For each hotel in the data set, a time series of binary observations was generated by instantiating a binary random number generator with the hotel's reputation score. From these observations, the trust estimate $t = \hat{p} = \frac{x}{n}$ can be computed at each point n in time in the time series, where x is the number of successes observed until point n. Obviously, $x \leqslant n$. The parameter c is computed as the *Wilson Interval Certainty Estimator* introduced in Chapter 3.1.5, Definition 12, p. 64. For each hotel, the supervised prediction result from the top performing random forest estimator (*regRF*, see, Table 12) was determined. Supervised predictions were generated using leave-one-out crossvaligdation.

The instantiation of parameter f was tested with four different values:

- $f = 0.5$: instiantiating f with its *CertainTrust* default value to provide a baseline;
- f as the supervised prediction result: instantiating f with the probability estimate to assess the impact feature-based prediction;
- f as the expectation value of the supervised prediction: instantiating f as the *CertainTrust* expectation value of the supervised prediction result, utilising the mapping in Definition 46;
- f as the mean of the training set: instantiating f as the average reputation of all hotels to determine the discriminative quality of the feature set.

Plotting the convergence of the trustworthiness estimation to the reputation scores of the hotels using different instantiations of f over time will reveal the actual informativity of the initial trust value. As time progresses, the effect of the initial trust value f diminishes, while early on when information is scarce, it dominates the prediction. Ideally, using supervised prediction methods in order to determine the initial trust value f yields an individual informative prior for each hotel. If the discriminative power of the feature set used for learning is sufficient, using the supervised prediction result to instantiate f should minimise the overall error and yield exact informative priors. However, as outlined above, the discriminative power of the features in the hotel data set is low. In the absence of discriminative features, the supervised learning machines used for the estimation tend to minimise the prediction error by approximating the mean value of the target scores in the training set.

In light of the low discriminative power of the features, it is to be expected that the mean absolute error for instantiating f as the output of the supervised estimator is highly similar to instantiating f with the mean value of the reputation scores in the training set. Similarly, instantiating f as the expectation value of the supervised prediction is expected to be similar to the default instantiation $f = 0.5$, as the certainty value used in the mapping in Definition 46 is utilising the NRMSE, which at 0.896 (Table 12, p. 200), leads to a low certainty value of only $c = 0.104$. Overall, a significant difference between the former two methods and the latter two can be expected, with the methods taking the distribution of the reputation scores into account (mean of reputation score in sample and the direct output of the supervised estimator) over the other two methods.

Figure 30 shows the overall effect of using different instantiations of the initial trust value f in *CertainTrust*. The figure depicts the mean absolute error between the *CertainTrust* expectation value and the reputation score averaged over all hotels in the data set. The expectations are largely confirmed, in that the instantiation of f with either the mean of the recommendation scores in the training set or the output of the supervised estimator significantly outperforming the other methods for instantiation (Wllcoxon test for significance [203], $p < 0.001$). A non-significant trend is observable that indicates that the instantiation of f based on the supervised estimator outperforms the instantiation with the mean of the training set; however, this advantage is minute.

The results are indicative of the low discriminative power of the feature set. The slight advantage of the supervised methods over the instantiation with the mean of the training shows that the sample of hotels in the data set, the observed features and the distribution of both regressand recommendation scores and regressor features are not yielding a significant advantage. Given the distribution of the recommendation scores for the hotels in Figure 28, p. 195 and Table 11, p. 194, the features are required to be highly discriminative, given the low number of hotels with a low recommendation score. Nonetheless, a marked performance advantage can be observed for instantiating f with an informative prior instead of a non-informative one.

This, however, does not mean that it is generally advisable to use the average recommendation score in the sample as an instantiation for f. As a matter of fact, a closer investigation of the impact of the overall distribution of the reputation scores in the sample on the achieved performance is warranted. To do so, the hotel data set was subdivided into quartiles according to recommendation scores (Figure 31, p. 211).

As can be seen in Figure 29b, p. 203, the predictive quality decreases almost linearly with its distance to the mean of the recommendation score. This is a clear indication that supervised estimator

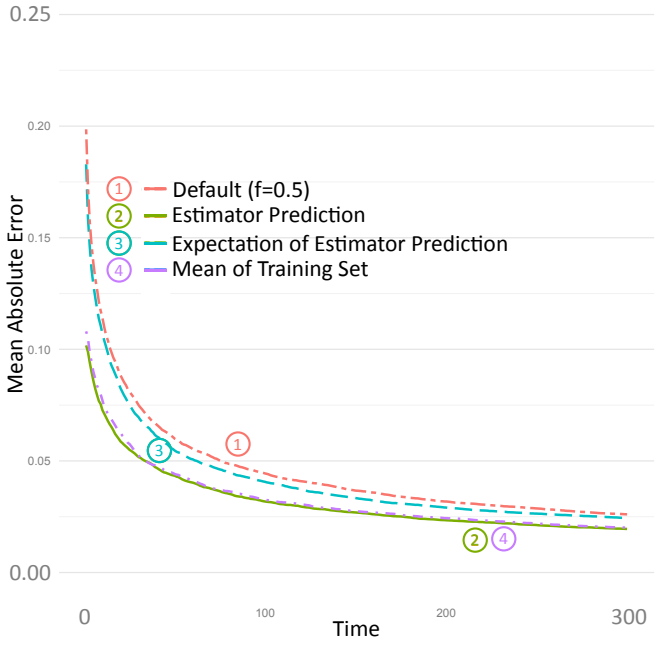
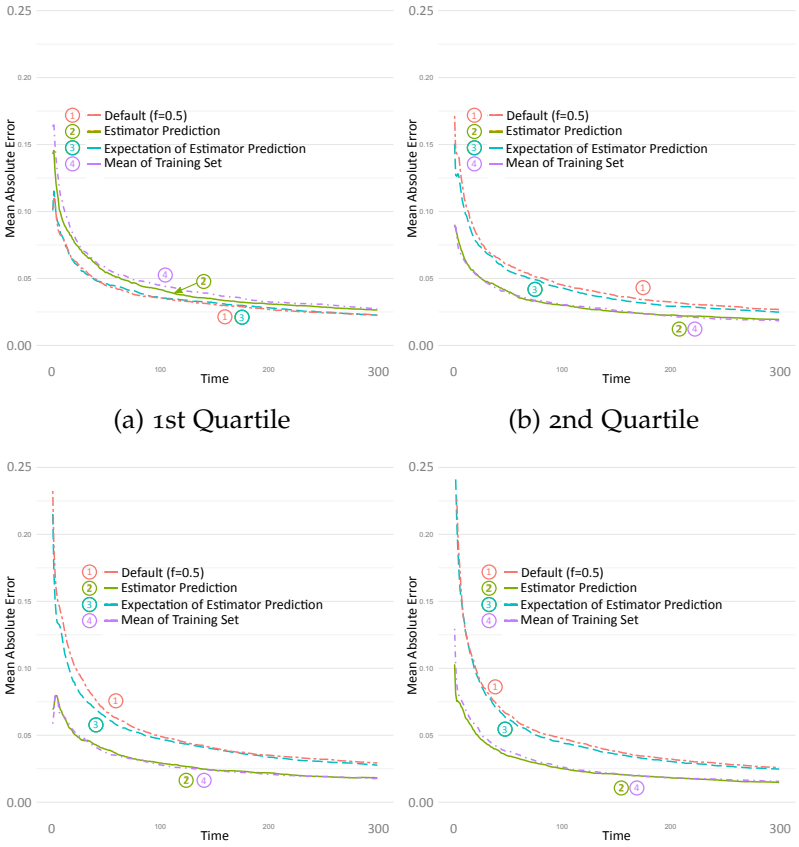


Figure 30: Impact of initial trust values f on predictive performance.

has minimised its error function by simply learning the mean value of the recommendation scores. This behaviour is expected to manifest itself in a poor predictive performance in the first quartile when instantiating f with the supervised estimator predictions. The second and third quartiles are clustered around the mean, while the fourth quartile is still reasonably close to the mean, when compared to first quartile (consult Table 11, p. 194).

As can be seen from Figure 31a, hotels in the first quartile are considerably overestimated with regard to their recommendation score when instantiating f with the supervised prediction. The resulting mean absolute error for hotels in the lowest quartile is significantly higher for instantiations with an informative than with a non-informative prior $f \approx 0.5$. This corroborates the expectation that the supervised estimator is unable to identify hotels with a low recommendation score, given the data set at hand.

Given the inability of the supervised estimator to correctly identify low scoring hotels, the initial assessment of hotels is biased towards the mean of the data set's recommendation scores. This leads to an overestimation of the recommendation score of low scoring hotels. This reveals problems in the general stereotyping approach, yet, it is also informative. Specifically, the application of supervised methods (and stereotyping approaches) requires careful analysis of the data set at the core of the model creation of the supervised learner. Without a representative data set, the prediction model will yield incorrect generalisations and consequently inaccurate estimates of the target variable.



(a) 1st Quartile (b) 2nd Quartile

(c) 3rd Quartile (d) 4th Quartile

Figure 31: Impact of initial trust values f on predictive performance (by quartile).

In the given data set, the regressand variable is clustered comparatively tightly around its mean value. The left-hand tail of its distribution is relatively light, with only a few instances below a value of 0.5. Thus, only very few hotels with bad recommendation scores exist for learning, while at the same time these hotels have comparably few recommendations that make up their low scores. This raises the questions whether or not the low ratings are actually justified, as the mean score for each of these has a comparatively high probability of being caused by sampling error (or even manipulation). As mentioned before, this behaviour likely stems from preferential attachment and selection processes that weed out bad performing hotels quickly. This effect is somewhat desirable in reputation systems, it can, however, limit the capabilities of such a system to yield a representative data set for supervised learning and stereotyping trust models. It is therefore important to ascertain the discriminative quality of the available features.

5.2.5.3 *General Observations on the Application of Supervised Learning to Reputation Data*

Determining whether or not the features in a particular data set requires a statistical analysis of the the and training the supervised estimator. While a simple analysis of the data set can establish the unsuitability of a data set for supervised learning in some simple cases, training an estimator with the data set (or a representative subset thereof) is required most of the time. Thus, determining the suitability of supervised methods in a specific scenario typically already yields the trained estimator.

Therefore, once the feasibility of supervised learning is to be checked, supervised estimators are – typically – already being trained. Depending on the quality of these trained estimators, three different possibilities for their use exist:

1. use as a standalone estimator in a stereotyping trust model, which is feasible only if the estimator has very good predictive performance;

2. use as a supplementary estimator in a feedback-based trust model, for instance by instantiating the initial trust value (i.e., an informative prior) with the estimator output;

3. discarding of the supervised estimator if its predictive performance is unsatisfactory.

While the quantification of what constitutes a *very good predictive performance* is subjective, very low error values are of course desirable. Clearly, with NRMSE values of close to 0.9, the estimators trained on the hotel data set do not perform particularly well. Given the range of the regressand, $[0;1]$, RMSE values of close to zero are necessary to consider a supervised estimator fit for use in stereotyping trust models.

For intermediate performance, the integration of supervised methods with (Bayesian) feedback-based models offers a way of leveraging feature-based supervised methods. The potential advantage over the standard instantiation $f = 0.5$ can be observed in Figure 30, p. 210. However, a closer examination of the prediction results is still necessary, even if overall predictive quality is good. In the hotel data set, it becomes obvious (Figure 31a) that the supervised predictor fails to correctly identify bad hotels. This can, of course, lead to unwanted selection and bad experiences, countering the elimination effect of the reputation system. However, since this quick weeding-out of under performers, which tends to be highly exploitative, as can be seen in Figure 28d, p. 195, results in bad hotels having only very few ratings, thereby increasing the chance of unfair elimination. From a systemic point of view, it might be desirable to keep those hotels in the selection process longer, in order to build a more representative data set

– even though doing so increases the risk of falsely selecting poorly in specific instances. This trade-off between system-wide increased predictive performance and loss due to poor selection presents an interesting avenue for future research.

A look at other data sets appears to indicate that data sets generated from reputation systems appear to have distributional properties that make their use in supervised prediction difficult. In particular, strong class imbalances, similar to those in the hotel data set, can be observed in a many real world reputation systems. Hu et al. [91] report asymmetric bi-modal distributions for many product review sites; for instance, the reputation system of the car-sharing site *BlaBlaCar*[5] generated a data set of approximately 190,000 ratings that, from a five-categorial rating scale, yielded 1.1 per cent of ratings in category *one* (lowest) and 98.9 per cent of ratings in category *five* (highest), with the intermediate categories being unrepresented[6]. A rating score distribution that is similar to the hotel data set can be found in the Netflix Prize data set of movie ratings[7], containing almost 500,000 members' ratings for about 18,000 movies[11]. In the five-categorial rating system, the lowest two categories account for only 4 and 9 per cent of the ratings, while categories *three*, *four* and *five* account for 28, 33, and 26 per cent, respectively. Apparently, effects are at play, which hint at the distribution observed in the hotel data set not to be a mere fluke. Rather, systemic effects might be culpable, for instance selection or confirmation bias. For the development of stereotyping approaches it is therefore a mandatory next step to investigate those effects on data set generation, and where possible compensate for their potentially negative influence on prediction quality.

5.2.6 *Section Summary*

In the previous section, the application of powerful supervised machine learning methods to a real-world data set was investigated. The goal of this was to determine, how and to what extent features provided for instance in a hotel booking and rating website can be used to enable generalised trust estimations. In a first step, the requirements for the application of supervised machine learning methods, so-called regression machines, to trustworthiness assessment were outlined. The impact, on a real-world dataset, of exploitative selection on the data generation process and how this affects predictive performance, was discussed. Subsequently, a mapping from estimator output to a belief logic representation that enables the use even of weak predictive results within the framework of trust assessment en-

5 http://www.blablacar.com

6 see also: http://tomslee.net/2013/09/some-obvious-things-about-internet-reputation-systems.html, retrieved August 2014.

7 http://www.netflixprize.com/

sembles was presented. Predictive results of the application of feature-based supervised methods within the framework of *CertainTrust* have been evaluated, as they have been generated from the hotel rating data set acquired for testing purposes. This data set shares properties with other data sets of user ratings generated by reputation systems, for instance, the Netflix Prize data set.

Using reputation systems in service selection, particularly when non-negligible resources are at stake, reinforces a trend towards exploitation. This has effects on the data that is generated and available for future trust assessment. The resulting complex adaptive system of trust assessment, selection and data generation merits closer attention in the future. For this, a data-centric, rather than a model-centric, approach to investigating the dynamics of trust and reputation systems is necessary. Developing flexible, component-based trust management approaches, standardised evaluation methodologies and a systematic collection and analysis of trust related datasets, in the form of a publicly available reference library for testing, are important next steps. In case of the hotel data set, the prediction result show that the set of selected features – even the data set as a whole – is of insufficient quality to guarantee the applicability of stereotyping schemes by themselves.

The particular distribution of the regressand values in the data set have a considerable impact on the ability of the supervised learners when building a model that can be used for stereotyping. Consequently, the straightforward application of relatively simple machine learners proposed in a number of stereotyping trust models (for instance, [32, 52, 92, 129]) is likely to be insufficient when faced with real-world data. Rather, significant effort has yet to be expended to make such models applicable, particularly with regard to data generation and the selection of meaningful features.

5.3 CHAPTER SUMMARY

In this chapter, possibilities for extending the generalisability of the trust model introduced in Chapter 3 and Chapter 4 by using feature-based supervised prediction were presented. For one, the dedicated modelling of particular roles and the trust delegation between them was shown to be principally possible as an extension to existing feedback-based trust models (Section 5.1, p. 162). For another, a more general approach for feature-based generalisation using model-free, supervised machine-learners, was introduced in Section 5.2, p. 183.

Generally speaking, the primary goal of the methods introduced and applied in this chapter is to imbue generalisability into feedback-based trust models. That is, to allow an estimation of trustworthiness of a trustee based on features said trustee exhibits and that can be observed by a truster. The basic principle applied is the delega-

tion of trust – with the exception of the insurance case in Section 5.1, which simply minimises risk. To facilitate the delegation, entities are grouped according to some role or set of features they possess. In Section 5.1, three such roles were defined, together with rules on how to transfer trust to the trustee: *certifiers*, *insurers* and *coalition partners*. The features that a trustee exhibits in the context of Section 5.1 is the association with a certifier, insurer or coalition partner. The delegation of trust in this scenario occurs from the certifying, insuring or coalition partners to the trustee, that is, from one group of entities to another, by the means of roles and rules hardcoded into the trust model. Similar rules can be created for delegation within a group of trustees, for instance, transferring trust among trustees sharing one or more specific features, such as geographic origin.

This is partially derived from the social practice of learning *stereotypes* and (pre-)judging or discriminating according to these. Once a stereotype is learned, it can be transferred from existing and known trustees to new and unknown ones, in the process transferring trust from an entity level to a more abstract and general plane. Hardcoding stereotypes into trust models is infeasible because of the potentially very large number of features a trustee can exhibit. Additionally, stereotypes can arise from combinations of features and the correlations between them, making explicit, human-driven modelling unpractical.

Section 5.2 uses non-parametric, model-free supervised machine learners to explore how such learners can be used for trustworthiness estimation by implicitly extracting stereotypes from a real-world data set. In a subsequent step, a mapping from the output of the machine learner to *CertainTrust* opinions is given that permits their flexible integration with feedback-based trust models. The difficulty of applying stereotyping methods, in the form of sophisticated supervised machine learners, to a data set generated from a real-world reputation system, is illustrated and the results are discussed. The relatively modest success achieved by the application of powerful machine learners raises the question to what extent stereotyping trust models are practicable when using reputation scores as a basis for their training. While related work on stereotyping trust models has focussed on modelling, the results obtained in this chapter from a data set that shares representative qualities with other data sets from reputation systems, present a new research direction for stereotyping trust models, to be tackled in future work: How model assumptions hold in the face of real-world data and how stereotyping can be usefully applied to reach generalisability and benefit the truster. Specific contributions in this chapter include:

- Trust model extensions to offer limited generalisability leveraging specific roles and relations that can be encountered in

e-commerce interactions. Three specific examples were chosen to show the principal practicability of such extensions:

  - *certifiers*, which certify the service quality, and hence the trustworthiness, of a certified trustee. A trust delegation mechanism is provided for partially transferring trust in a certifier onto the certified trustee.
  - *insurers*, which provide assurance against loss potentially incurred from an untrustworthy trustee. A trust delegation mechanism is provided that influences decision trust, that is, the expected utility of an interaction with the insured trustee.
  - *coalition partners*, which are associated with the trustee in a (semi-)permanent fashion. A trust delegation mechanism is provided for partially transferring trust in coalition partners onto the trustee partner in a coalition.

- Application and evaluation of powerful supervised machine learning approaches to a real-world data set with a regressand value generated from a reputation system. The distribution of the regressand value follows a distribution that is both typical of those from a reputation system and is adverse to the successful application of supervised methods. The predictive results suggest that the model-centric approach taken in the design of existing stereotyping trust models needs to be complemented by a data-centric analysis and that idealised simulations are insufficient to ascertain feasibility.

- A mapping from the output of supervised machine learners to *CertainTrust* opinions, enabling the integration of supervised learning with feedback-based trust models. Thereby, generalisable information contained in the features of a given data set can be harnessed, even if the prediction quality is only mediocre. The prediction of the estimator is mapped directly to the *CertainTrust* trust parameter t, while a statistical measure of the prediction quality – in this case, the *normalised root mean squared error* (NRMSE) – is mapped to the certainty parameter c.

The goal of introducing the advances discussed in this chapter was to overcome negative aspects inherent to purely feedback-based trust and reputation system, namely, newcomer discrimination, sole reliance on potentially scarce feedback information and foregoing of information encoded in observable features. By hardcoding indicators of trustworthiness, specific features of interactions can be explicitly modelled and introduced into trust-based decision making. The three presented indicators (certificates, insurance, and coalitions) can, in the future, be expanded into a toolset of trust-building extensions for feedback-based trust and reputation systems. By investigating the

application of supervised machine learners to a real-world dataset and providing a mapping to *CertainTrust* opinions, a flexible way of integrating feature-based stereotyping and feedback-based trust models was established. However, one key finding when applying supervised learners was their dependence on representative and discriminatory feature sets, that may not always be available. Thus, stereotyping, as it has been proposed in the literature, may not be as effective in real world applications as it appears in simulations. By integrating stereotyping into the computation of the initial trust value parameter f of the *CertainTrust* model, a way was provided, however, to also leverage feature sets that are only weakly discriminatory. As a result, the strengths of feedback-based and feature-based/stereotyping approaches can now be combined.

Overall, in this chapter methods for improving the generalisability of feedback-based trust models have been proposed. From hardcoding indicators of trustworthiness, the approach has been extended to flexibly include stereotype-like results from non-parametric, model-free supervised learners. Particularly the results gathered from the application of the latter points at a further need for researching trust models not just from a model-centric, but rather also from a data-centric point of view.

6 CONCLUSION AND OUTLOOK

Over the course of the previous chapters, a trust model was sequentially developed (and evaluated) from a core trust estimation model, built upon point and interval estimation techniques for binomial and multinomial proportions. The core estimation model was augmented with methods for trust propagation, the combination of trust sources and a mapping for the integration of stereotype-based supervised estimators. This chapter summarises the main contributions and findings of this thesis and presents an outlook.

6.1 CONCLUSION

This thesis contributed advanced methods for trustworthiness estimation that in their entirety constitute the *Multinomial CertainTrust* model, a considerable extension of the binomial *CertainTrust* model introduced by Ries [173]. The statistics of trustworthiness estimation were given a particular focus, by first explicating the assumptions behind the trust and trustworthiness estimation in Chapter 2.1.3. Next, in the design of the core estimation model in Chapter 3, appropriate statistical methods were used to improve and extend the state of the art in trustworthiness estimation, both in the binomial and multinomial case. The focus on statistically sound methods is maintained in Chapter 4, where hypothesis testing leveraging exact test methods is introduced to support trust propagation and to cope with non-stationary. Finally, in Chapter 5, the integration of methods for generalising from features and stereotypes in trustworthiness estimation is discussed.Thus, a complete trust model for binomial and multinomial feedback is presented that can also integrate stereotype-based trustworthiness predictions.

In particular, the *Multinomial CertainTrust* model meets the requirements postulated in Chapter 2.2.1:

- Requirement 1 (Probabilistic Computation, Representation and Interpretation of Trust and Certainty): *Multinomial CertrainTrust* extends state-of-the-art models (for instance, [107, 173, 194]) by offering a complete probabilistic trustworthiness estimation model that extends the foundations of Bayesian estimation from just the trustworthiness estimate to the computation of certainty scores. Both trust and certainty are computed using estimators that harness the Bayesian posterior distribution of the trustworthiness estimate. The methods leveraged to achieve statistically well-founded certainty estimation are based on interval esti-

mation techniques; while Teacy' in *TRAVOS* [189] and Wang & Singh [197] also base their certainty estimators on properties of the posterior distribution of the trustworthiness estimate. However, the certainty estimators of the *Multinomial CertainTrust* model, by being derived from interval-based estimators that model the potential dispersion of the trustworthiness estimate, are by design capable of being extended to the multinomial case – a capability not shared by either *TRAVOS* or Wang & Singh's model.

- Requirement 2 (Binomial and Multinomial Estimation Model): *Multinomial CertainTrust* is capable of handling both binary and, more generally, m-categorical feedback, $m \in \mathbb{N}, m \geqslant 2$. The estimation model provides estimation techniques for binomially and multinomially distributed feedback that extend to both trustworthiness and certainty estimation. The Bayesian methods used, relying on Beta and Dirichlet posteriors, are applied and adapted, guaranteeing adherence to this requirement by building on established statistical methods.

- Requirement 3 (Trust Propagation and Combining Trust Sources): By extending discounting, consensus and fusion operations from the original *CertainTrust* [173] and *CertainLogic* [77, 175] to the multinomial case, *Multinomial CertainTrust* has capabilities for trust propagation and combining trust sources that are similar in functionality to *Subjective Logic* [103]. By also providing advanced fusion methods, that is, weighted and conflict-aware fusion, for both binomial and multinomial models, progress over the start of the art is achieved.

- Requirement 4 (Integration of Non-Frequentist Information): *Multinomial CertainTrust* has at its heart an estimation model founded upon Bayesian statistics. The Bayesian prior allows for a straightforward integration of subjective, non-frequentist information. Additionally, the use of fusion operations also provided in this thesis over another avenue for the integration of information. An example for this is the use of stereotyping-derived trustworthiness estimates that can be integrated either in the prior or via a mapping presented in Chapter 5.2.

- Requirement 5 (Changing Trustee Behaviour): The standard approach to dealing with changing behaviour, in the form of changes in the unobservable true trustworthiness of a trustee, is by applying ageing or fading to the collected evidence, thereby diminishing the impact of older information and favouring newer evidence. Ageing is supported by *Multinomial CertainTrust*. However, ageing puts a limit on the achievable accuracy of the trustworthiness estimate, thus bounding the certainty estimate. In or-

der to deal with this, change point detection was introduced to the field of trustworthiness estimation and shown to be effective in discovering behavioural changes. When combined with conservative ageing, it increases the responsiveness to change exhibited by the estimation model significantly. This gives *Multinomial CertainTrust* advanced capabilities to deal with non-stationarity of the estimand parameter.

By fulfilling the listed requirements, *Multinomial CertainTrust* provides a comprehensive model for trustworthiness estimation from binary or m-categorical evidence. The estimation model introduced in Chapter 3 is built on a foundation of well-proven methods from the field of statistics, permitting a probabilistic interpretation of all estimates the model provides. In its multinomial form, *Multinomial CertainTrust* is, to the best of our knowledge, the only evidence-based trust model capable of providing statistically-derived (from a Bayesian Dirichlet prior) certainty estimates for each of its $m > 2$ estimates in a multinomial, m-categorical setting. Other models either only support the binomial case of trustworthiness estimation, which does not allow for adjustments in the granularity of the desired feedback categories when deploying the model as part of a real-world trust or reputation system, or do not offer sophisticated certainty estimation methods. Among the former are advanced binomial trust models like *TRAVOS* [189] and Wang & Singh's model [197], the latter category is epitomised primarily by the otherwise comprehensive *Subjective Logic* [103]. Most models, however, offer neither properly statistically-derived certainty estimates nor multinomial trustworthiness estimation.

Beyond the estimation model, Chapter 4 provides additional functionality for processing trust-relevant information. The most basic functionality is the support for trust propagation, a staple of trust models in the state of the art. Standard methods and robustness improvements from the original *CertainTrust* were retained and expanded to be applicable to the multinomial case. The state of the art was advanced by exact methods for comparing distributional information to benefit the assessment of the reliability of recommenders and change point detection for detecting non-stationarity in trustee behaviour. In both applications, exact test methods compared favourably to the state-of-the-art.

Considering trust models with a closer look at the statistical methods that are and can be used in their design, has revealed gaps in the state-of-art, but has at the same time provided the tools for developing methods to close these gaps. The work that has been presented and evaluated in this thesis is considered to be valuable and significant contributions in field of evidence-based trust models.

In Chapter 5, methods for extending trustworthiness estimation beyond the purely Bayesian approach were addressed. The first part on

hardcoding indicators of trustworthiness into evidence-based trust models represented a proof of concept and illustrates the general feasibility of indicators that are not directly linked to the past performance of a particular trustee. The second part of Chapter 5 addresses the use of supervised learning in stereotyping trust models and how they can be integrated into evidence-based models. Here it has been shown that the assumptions regarding the distributions of both the trust estimates and the features, as well as the discriminatory power of these features, made in state-of-the-art stereotyping models [31, 52], tend to be too optimistic. Testing of powerful machine learners, such as various *Decision Trees* and *Random Forests*, against real-world data revealed that, while some information can be extracted, stereotyping alone proves insufficient and offers only marginal improvements.This suggests that stereotyping approaches serve best in a supporting role to evidence-based models. Most importantly, however, the findings of Chapter 5.2 stress the need to look at trustworthiness estimation not only from a modelling perspective, i.e., by discussing methods that can be applied to the problem and then verifying the data in ideal simulations, but also from a data perspective that considers the data that is used for inference by (stereotyping) trust models.

6.2 OUTLOOK

This thesis has provided a comprehensive trust model capable of handling binomial and multinomial feedback, using advanced methods for trustworthiness estimation and the processing of trust-relevant information. Nonetheless, numerous aspects for further research still remain.

One of the assumptions made for the estimation model presented in this thesis is categorical feedback. This assumption is made in order to apply Bayesian statistics and estimators that have a well-defined behaviour. By demanding categorical feedback, Beta and Dirichlet priors can be enforced, considerably simplifying the process of estimating trustworthiness and interpreting it in a probabilistic manner. Continuous feedback, however, might be preferable in some applications. Here, distributional assumptions cannot be enforced so easily and the ready accessibility of conjugate priors may not be given. Rather, the prior distribution itself would have to be estimated. The related work that uses continuous ratings, such as *FIRE* [93, 94], simply assumes Gaussianity of the underlying distribution. An assumption that may not be warranted. This leaves room for future research on both the distribution of continuous feedback in a number of application areas and estimation methods fit for modelling such feedback.

In Chapter 5.1, expected utility theory was briefly discussed as a decision making tool. The integration of trustworthiness estimation and utility, as well as general decision theoretic aspects of trust and

trusting choices deserve closer investigation. This can extend from a system model to determining user behaviour under the influence of a trust and reputation system.

DATA-CENTRIC APPROACH TO TRUST AND REPUTATION SYSTEMS : Aside from decision theoretic aspects, the feedback behaviour of users of trust and reputation systems leads to specific patterns in the data generated by these systems (see [91]). Various social biases might be responsible for generating specific distributions of trust and reputation scores, as may be the role that trust and reputation systems play to weed out badly behaving trustees, as a soft security mechanism. The distribution of the regressand trust and reputation scores can considerably affect the performance of supervised, stereotyping trust models. For instance, in some application fields, feedback may only be given in case of trustee defection, resulting in overwhelming negative feedback. Conversely, in well-governed (social) systems, regulation may enforce quality standards that are generally followed, resulting in a majority of positive feedback. Either case impacts the predictive quality of a trust model. Therefore, a data-centric investigation of trust and reputation systems and trustworthiness estimation techniques is warranted.

TRUST CONTEXT AND TRANSFER BETWEEN CONTEXTS :Another aspect that was not explicitly addressed in this thesis is the notion of context. Trust is dependent on context and trustees may behave differently in different contexts, depending on their abilities and agendas; the capabilities of trusters to evaluate their interactions with trustees may also vary from context to context. Thus, the impact of context on trust formation is an interesting area of future research, as is the transfer of trust between contexts.

PRIVACY-ENABLED TRUST AND REPUTATION SYSTEMS : As is evident from the methods for collecting trust-relevant information, that is, recording past behaviour and correlating observable features with behaviour, trustworthiness estimation may be privacy-invasive. Privacy-friendly trustworthiness estimation is still a little researched field. Developing technologies that allow a user to demonstrate its trustworthiness in a privacy-friendly manner would help to curb the desire to collect evermore data to increase the certainty of a trustee's good intentions. Combining approaches for a minimisation of data revealed and used for trust estimation with accurate trustworthiness estimators promises to be an interesting field for future research.

INTEGRATION OF CERTIFICATE-BASED TRUST AND COMPUTATIONAL TRUST : The combination of trusted computing and computational trust still remains a field with considerable room for research. Inte-

grating probabilistic approaches with certificate-based security mechanisms may cure at least some of the problems encountered in todays internet security infrastructure, such as the Web-PKI. Establishing the trustworthiness of certificates is a challenging task that has room for future work, as is the computation of a trustee's trustworthiness from certificates it can provide. Similarly, when demanding trustworthy systems, computational trust and trusted computing methods may be integrated fruitfully. Both approaches can, conceivably, complement each other when determining and guaranteeing a (minimum) level of trustworthiness required for a particular (inter)action.

COMPUTATIONAL TRUST FOR SECURITY : Aside from integrating security mechanisms and computational trust to evaluate and guarantee the trustworthiness of systems, computational trust can also be applied to determine the trustworthiness of particular security mechanisms. This permits not only the assignment of trustworthiness estimates to, for instance, an encryption method for establishing a secure communication channel, but also gives an insight into how crucial this specific mechanism is to guarantee a trustworthy end-to-end interaction.

These and numerous other challenges remain in the field of computational trust modelling. With the increasing proliferation of internet access and the speed at which the real and the virtual converge, building and estimating trust in this emerging environment will remain a challenging task in the future.

APPENDIX

A

DEFINITIONS OF TRUST

Trust has been researched by a multitude of researchers active in different fields of study. Due to the broad manner in which the term trust can be interpreted and the wide acceptance of the concept's positive effects, such fields range from the social sciences (in particular history, psychology, sociology) and economics to computer science and engineering. The following listing, without claiming completeness, comprises several popular definitions of the term trust by scholars from these diverse disciplines.

> *Trust is an important lubricant of a social system. It is extremely efficient; it saves a lot of trouble to have a fair degree of reliance on other people's word.*
>
> – Arrow [7]

> [Trust is] *a state involving confident positive expectations about another's motives with respect to oneself in situations entailing risk.*
>
> – Boon and Holmes [21]

> *On-line trust is an attitude of confident expectation in an online situation of risk that one's vulnerabilities will not be exploited.*
>
> – Corritore, Kracher and Wiedenbeck [37]

> [Trust is] *the decision to rely on another party (i.e. person, group or organization) under a condition of risk.*
>
> – Currall and Inkpen [39]

> *Trust is the confidence that one will find what is desired from another rather than what is feared.*
>
> – Deutsch [44]

> *Trust is a normative notion in the sense that an essential ingredient in all cases of trust and trustworthiness is the existence of a set of norms that provide the motivation to cooperate.*
>
> – Elgesem [51]

Trust is the expectation that arises within a community of regular, honest, and cooperative behavior, based on commonly shared norms, on the part of other members of that community.

– Fukuyama [61]

Trust (or, symmetrically, distrust) is a particular level of the subjective probability with which an agent assesses that another agent or group of agents will perform a particular action, both before he can monitor such an action (or independently of his capacity ever to be able to monitor it) and in a context in which it affects his own action.

– Gambetta [64]

Trust is a qualified belief by a trustor with respect to the competence, honesty, security and dependability of a trustee within a special context.

– Grandison and Sloman [73]

[Trust is the] *undertaking of a risky course of action on the confident expectation that all persons involved in the action will act competently and dutifully.*

– Lewis and Weigert [127]

[Trust is] *an effective form of complexity reduction.*

– Luhmann [132]

Trust, in general, is taken as the belief (or measure thereof) that a person (the trustee) will act in the best interest of another (the truster) in a given situation, even when controls are unavailable and it may not be in the trustee's best interest to do so.

– Marsh and Dibben [136]

[Trust is] *the willingness of a party to be vulnerable to the actions of another party based on the expectation that the other will perform a particular action important to the trustor, irrespective of the ability to monitor or control that other party.*

– Mayer, Davis and Schoorman [138]

Trust is a willingness to be vulnerable based on the expectation that the other party is reliable, open, competent and compassionate.

– Mishra [147]

Trust is a subjective expectation an agent has about another agent's future behavior.

– Mui [151]

To trust is to accept or neglect the possibility that things will go wrong. To have trust in the narrow sense, or "real" intentional trust, is to accept or neglect the possibility that a partner will utilize opportunities for opportunism even if it is in his interest to do so.

– Nooteboom [158]

Trust is a bet about the future contingent actions of others.

– Sztompka [188]

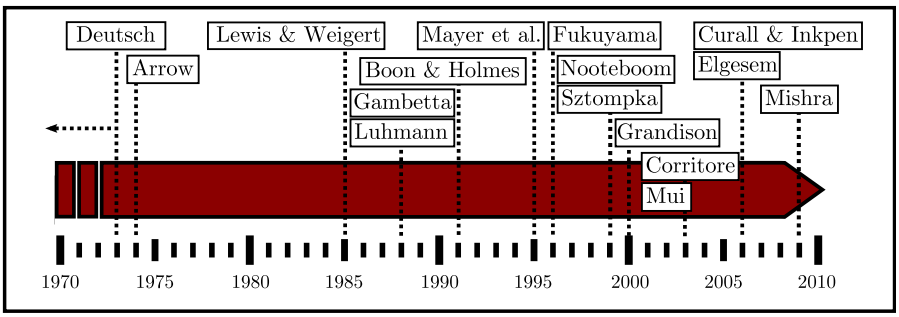


Figure 32: Trust Definitions, Timeline

B

PRE-COMPUTED TABLES FOR $C_{J;(100\cdot z)\%}(x, n)$

x	**n= 2**	**n= 3**	**n= 4**	**n= 5**	**n= 6**	**n= 7**
0	0.333	0.465	0.555	0.621	0.67	0.708
1	0.122	0.215	0.312	0.394	0.461	0.515
2	0.333	0.215	0.246	0.304	0.363	0.417
3	—	0.465	0.312	0.304	0.334	0.373
4	—	—	0.555	0.394	0.363	0.373
x	**n= 8**	**n= 9**	**n= 10**	**n= 11**	**n= 12**	**n= 13**
0	0.738	0.762	0.783	0.8	0.815	0.827
1	0.56	0.598	0.63	0.657	0.681	0.701
2	0.464	0.505	0.541	0.573	0.6	0.624
3	0.414	0.452	0.487	0.518	0.547	0.572
4	0.398	0.427	0.457	0.485	0.512	0.537
5	0.414	0.427	0.447	0.47	0.492	0.515
6	0.464	0.452	0.457	0.47	0.486	0.504
7	0.56	0.505	0.487	0.485	0.492	0.504
x	**n= 14**	**n= 15**	**n= 16**	**n= 17**	**n= 18**	**n= 19**
0	0.838	0.848	0.857	0.865	0.871	0.878
1	0.719	0.736	0.75	0.763	0.774	0.785
2	0.646	0.665	0.683	0.698	0.713	0.726
3	0.595	0.616	0.635	0.652	0.668	0.682
4	0.56	0.581	0.6	0.618	0.634	0.649
5	0.536	0.556	0.575	0.592	0.609	0.624
6	0.522	0.54	0.558	0.574	0.59	0.605
7	0.518	0.533	0.548	0.563	0.577	0.591
8	0.522	0.533	0.544	0.557	0.57	0.582
9	0.536	0.54	0.548	0.557	0.567	0.578
10	0.56	0.556	0.558	0.563	0.57	0.578
x	**n= 20**	**n= 21**	**n= 22**	**n= 23**	**n= 24**	**n= 25**
0	0.883	0.889	0.893	0.898	0.902	0.905
1	0.795	0.803	0.812	0.819	0.826	0.832
2	0.738	0.748	0.758	0.768	0.776	0.784
3	0.696	0.708	0.719	0.729	0.739	0.748

4	0.663	0.676	0.688	0.699	0.71	0.72
5	0.638	0.651	0.664	0.675	0.686	0.696
6	0.619	0.632	0.644	0.656	0.667	0.678
7	0.605	0.617	0.629	0.641	0.652	0.662
8	0.595	0.607	0.618	0.629	0.64	0.65
9	0.589	0.6	0.61	0.621	0.631	0.64
10	0.587	0.596	0.605	0.615	0.624	0.633
11	0.589	0.596	0.604	0.612	0.62	0.629
12	0.595	0.6	0.605	0.612	0.619	0.626
13	0.605	0.607	0.61	0.615	0.62	0.626
x	**n= 26**	**n= 27**	**n= 28**	**n= 29**	**n= 30**	**n= 31**
0	0.909	0.912	0.915	0.918	0.92	0.923
1	0.838	0.844	0.849	0.854	0.858	0.862
2	0.792	0.799	0.805	0.811	0.817	0.822
3	0.757	0.765	0.772	0.779	0.786	0.792
4	0.729	0.737	0.745	0.753	0.76	0.767
5	0.706	0.715	0.724	0.732	0.739	0.746
6	0.687	0.697	0.705	0.714	0.721	0.729
7	0.672	0.681	0.69	0.699	0.707	0.714
8	0.659	0.669	0.678	0.686	0.694	0.702
9	0.65	0.659	0.667	0.676	0.683	0.691
10	0.642	0.651	0.659	0.667	0.675	0.682
11	0.637	0.645	0.653	0.66	0.668	0.675
12	0.634	0.641	0.648	0.656	0.663	0.67
13	0.633	0.639	0.646	0.652	0.659	0.665
14	0.634	0.639	0.645	0.651	0.657	0.663
15	0.637	0.641	0.646	0.651	0.656	0.661
16	0.642	0.645	0.648	0.652	0.657	0.661
x	**n= 32**	**n= 33**	**n= 34**	**n= 35**	**n= 36**	**n= 37**
0	0.925	0.927	0.929	0.931	0.933	0.935
1	0.866	0.87	0.874	0.877	0.88	0.883
2	0.827	0.832	0.837	0.841	0.845	0.849
3	0.798	0.803	0.808	0.813	0.818	0.823
4	0.773	0.779	0.785	0.791	0.796	0.801
5	0.753	0.76	0.766	0.772	0.777	0.782
6	0.736	0.743	0.749	0.755	0.761	0.767
7	0.721	0.728	0.735	0.741	0.747	0.753
8	0.709	0.716	0.723	0.729	0.735	0.741

9	0.698	0.705	0.712	0.718	0.725	0.731
10	0.689	0.696	0.703	0.709	0.716	0.722
11	0.682	0.689	0.695	0.702	0.708	0.714
12	0.676	0.683	0.689	0.695	0.701	0.707
13	0.672	0.678	0.684	0.69	0.696	0.702
14	0.669	0.675	0.68	0.686	0.692	0.697
15	0.667	0.672	0.678	0.683	0.688	0.693
16	0.666	0.671	0.676	0.681	0.686	0.691
17	0.667	0.671	0.675	0.68	0.684	0.689
18	0.669	0.672	0.676	0.68	0.684	0.688
19	0.672	0.675	0.678	0.681	0.684	0.688
x	**n= 38**	**n= 39**	**n= 40**	**n= 41**	**n= 42**	**n= 43**
0	0.936	0.938	0.94	0.941	0.942	0.944
1	0.886	0.889	0.892	0.894	0.897	0.899
2	0.853	0.856	0.86	0.863	0.866	0.869
3	0.827	0.831	0.835	0.838	0.842	0.845
4	0.805	0.81	0.814	0.818	0.822	0.826
5	0.787	0.792	0.797	0.801	0.805	0.81
6	0.772	0.777	0.782	0.786	0.791	0.795
7	0.758	0.764	0.769	0.774	0.778	0.783
8	0.747	0.752	0.757	0.762	0.767	0.772
9	0.736	0.742	0.747	0.752	0.757	0.762
10	0.727	0.733	0.738	0.743	0.748	0.753
11	0.719	0.725	0.73	0.735	0.74	0.745
12	0.713	0.718	0.723	0.729	0.733	0.738
13	0.707	0.712	0.718	0.723	0.727	0.732
14	0.702	0.708	0.713	0.718	0.722	0.727
15	0.699	0.704	0.708	0.713	0.718	0.722
16	0.696	0.7	0.705	0.71	0.714	0.719
17	0.694	0.698	0.702	0.707	0.711	0.716
18	0.692	0.696	0.701	0.705	0.709	0.713
19	0.692	0.696	0.7	0.703	0.707	0.711
20	0.692	0.696	0.699	0.703	0.706	0.71
21	0.694	0.696	0.7	0.703	0.706	0.709
22	0.696	0.698	0.701	0.703	0.706	0.709

Table 17: Certainty from 1-Width of the 95 per cent Jeffreys prior interval.

C

AUXILIARY PROOFS

Proof for $E(\text{Beta}(\hat{p} \cdot n + \alpha_0, (1-\hat{p}) \cdot n + \beta_0)) = E(t, c, f)$.

$$
\begin{aligned}
& E(\text{Beta}(\hat{p} \cdot n + \alpha_0, (1-\hat{p}) \cdot n + \beta_0)) \\
&= \frac{\hat{p} \cdot n + \alpha_0}{\hat{p} \cdot n + (1-\hat{p}) \cdot n + \alpha_0 + \beta_0} \\
&= \frac{x + \alpha_0}{n + \alpha_0 + \beta_0} && \| \ \hat{p} = \tfrac{x}{n} \\
&= \frac{x + f \cdot (1 - C(n,\hat{p})) \cdot \frac{n}{C(n,\hat{p})}}{n + (1 - C(n,\hat{p})) \cdot \frac{n}{C(n,\hat{p})}} && \| \ \text{Def. 15, p. 70} \\
&= \frac{C(n,\hat{p}) \cdot x + C(n,\hat{p}) \cdot f \cdot (1 - C(n,\hat{p})) \cdot \frac{n}{C(n,\hat{p})}}{n} && \| \ \text{Expand fraction by } C(n,\hat{p}) \\
&= \frac{C(n,\hat{p}) \cdot x + f \cdot (1 - C(n,\hat{p})) \cdot n}{n} \\
&= C(n,\hat{p}) \cdot \tfrac{x}{n} + f \cdot (1 - C(n,\hat{p})) \\
&= C(n,\hat{p}) \cdot t + (1 - C(n,\hat{p})) \cdot f && \| \ \tfrac{x}{n} = \hat{p} = t \\
&= E(t, c, f)
\end{aligned}
$$

□

Proof of the Aggregation Property of Dirichlet Distributions (reproduced from [60]). Any Dirichlet distribution can be represented as a normalised set of Gamma distributed random variables. The Gamma distribution $\Gamma(\kappa, \theta)$ is given by the probability density function

$$f(x; \kappa, \theta) = x^{\kappa-1} \cdot \frac{e^{-\frac{x}{\theta}}}{\theta^{\kappa} \cdot \Gamma(\kappa)}$$

with shape parameter $\kappa > 0$, scale parameter $\theta > 0$ and $\Gamma(y) = \int_0^\infty t^{y-1} \cdot e^{-t} dt$. For the Gamma distribution the following property holds: If $X_i \sim \Gamma(\kappa_i, \theta)$ for $i = 1, 2, \ldots, n$ are independent Gamma distributed random variables of the same scale but different shapes. Then $S = \sum_{i=1}^{n} X_i \sim \Gamma(\sum_{i=1}^{n} \kappa_i, \theta)$. Let $z_i =\sim \Gamma(\alpha_i, 1)$, $q_i = \frac{z_i}{\sum_{i=1}^{k} z_i}$ for $i = 1, 2, \ldots, k$.

First, it has to be shown that $(q_1, \ldots, q_k) \sim \mathrm{Dir}(\alpha_1, \ldots, \alpha_k)$. Consider$\{z_i\}_1^k$ be given, and consider $z, q_1, \ldots, q_{k-1}$ as new variables. We relate them using the transformation T:

$$(z_1, \ldots, z_k) = T(z, q_1, \ldots, q_{k-1}) = \left(z \cdot q_1, \ldots, z \cdot q_{k-1}, z \cdot \left(1 - \sum_{i=1}^{k-1} q_i \right) \right)$$

The Jacobian matrix, i.e., the matrix of the first derivatives, of this transformation is :

$$J(T) = \begin{pmatrix} q_1 & z & 0 & 0 & \cdots & 0 \\ q_2 & 0 & z & 0 & \cdots & 0 \\ q_3 & 0 & 0 & z & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & & \\ q_{k-1} & 0 & 0 & 0 & \cdots & z \\ 1 - \sum_1^{k-1} q_i & -z & -z & -z & \cdots & -z \end{pmatrix}$$

$J(T)$ has determinant Z^{k-1}.

The change-of-variables formula gives the density f of $(z, q_1, \ldots, q_{k-1})$ as $f = g \circ T \times |\det(T)|$, where the joint density g of the original, independent random variables is given by

$$g(z_1, z_2, \ldots, z_k; \alpha_1, \ldots, \alpha_k) = \prod_{i=1}^{k} z_i^{\alpha_i - 1} \frac{e^{-z_i}}{\Gamma(\alpha_i)}$$

Substituting this into the change-of-variables formula yields

$$f(z, q_1, \ldots, q_{k-1})$$

$$= \left(\prod_{i=1}^{k-1} (z \cdot q_i)^{\alpha_i - 1} \frac{e^{-z q_i}}{\Gamma(\alpha_i)} \right) \left[\left(z \left(1 - \sum_1^{k-1} q_i \right) \right)^{\alpha_k - 1} \frac{e^{-z(1 - \sum_1^{k-1} q_i)}}{\Gamma(\alpha_k)} \right] z^{k-1}$$

$$= \frac{\left(\prod_{i=1}^{k-1} q_i^{\alpha_i - 1} \right) \left(1 - \sum_{i=1}^{k-1} q_i \right)}{\prod_{i=1}^{k} \Gamma(\alpha_i)} z^{(\sum_{i=1}^{k} \alpha_i) - 1} e^{-z}$$

Integrating over z yields the marginal distribution of $\{q_i\}_{i=1}^{k-1}$

$$f(q_1, \ldots, q_{k-1})$$

$$= \int_0^\infty f(z, q_1, \ldots, q_{k-1}) dz$$

$$= \frac{\left(\prod_{i=1}^{k-1} q_i^{\alpha_i - 1} \right) \left(1 - \sum_{i=1}^{k-1} q_i \right)}{\prod_{i=1}^{k} \Gamma(\alpha_i)} \int_0^\infty z^{(\sum_{i=1}^{k} \alpha_i) - 1} e^{-z}$$

$$= \frac{\Gamma(\sum_{i=1}^{k} \alpha_i)}{\prod_{i=1}^{k} \Gamma(\alpha_i)} \left(\prod_{i=1}^{k-1} q_i^{\alpha_i - 1} \right) \left(1 - \sum_{i=1}^{k-1} q_i \right)^{\alpha_k - 1}$$

which is a Dirichlet density. Thus, a normalised set of Gamma distributed random variables represents a Dirichlet distribution.

$$f(q_1, \ldots, q_{k-1})$$

$$= \int_0^\infty f(z, q_1, \ldots, q_{k-1}) dz$$

$$= \frac{\left(\prod_{i=1}^{k-1} q_i^{\alpha_i - 1}\right)\left(1 - \sum_{i=1}^{k-1} q_i\right)}{\prod_{i=1}^{k} \Gamma(\alpha_i)} \int_0^\infty z^{(\sum_{i=1}^{k} \alpha_i) - 1} e^{-z}$$

$$= \frac{\Gamma(\sum_{i=1}^{k} \alpha_i)}{\prod_{i=1}^{k} \Gamma(alpha_i)} \left(\prod_{i=1}^{k-1} q_i \ddot{o} \alpha_i - 1\right) \left(1 - \sum_{i=1}^{k-1} q_i\right)^{\alpha_k - 1}$$

We can now proof the aggregation property of the Dirichlet distribution using this and relying on the property of the Gamma distribution that if $X_i \sim \Gamma(\kappa_i, \theta)$ for $i = 1, 2, \ldots, n$ are independent Gamma distributed random variables of the same scale but different shapes, it follows that $S = \sum_{i=1}^{n} X_i \sim \Gamma(\sum_{i=1}^{n} \kappa_i, \theta)$.

Suppose $(Q_1, Q_2, \ldots, Q_k) \sim Dir(\alpha_1, \ldots, \alpha_k)$. Then we know that $Q_i = \frac{Z_i}{\sum_{i=1}^{k} Z_i}$, where $Z_i \sim \Gamma(\alpha_i, \theta)$ are independent. Let $\{A_1, A_2, \ldots, A_r\}$ be a non-trivial partition of $\{1, 2, \ldots, k\}$. Then,

$$\left(\sum_{i \in A_1} Q_i, \sum_{i \in A_2} Q_i, \ldots, \sum_{i \in A_r} Q_i\right)$$

$$= \frac{1}{\sum_{i=1}^{k} Z_i} \left(\sum_{i \in A_1} Z_i, \sum_{i \in A_2} Z_i, \ldots, \sum_{i \in A_r} Z_i\right)$$

$$\sim \frac{1}{\sum_{i=1}^{k} \Gamma(\alpha_i, 1)} \left(\Gamma(\sum_{i \in A_1} \alpha_i, 1), \Gamma(\sum_{i \in A_2} \alpha_i, 1), \ldots, \Gamma(\sum_{i \in A_r} \alpha_i, 1)\right)$$

$$\sim Dir\left(\sum_{i \in A_1} \alpha_i, \sum_{i \in A_2} \alpha_i, \ldots, \sum_{i \in A_r} \alpha_i\right)$$

□

D

SYBIL RESISTANT CONSENSUS

In [173], Ries presented a Sybil resistant extension to the consensus operator for binomial opinions. This extension is reproduced in the following and extended to the general multinomial case. Ries' approach introduces two additional parameters in the consensus operation that limit the maximum influence a single recommender (a single recommendation/opinion) can have on the resulting composite opinion:

1. *Normalisation parameter* N_R: When the total number of observations reported in a recommendation exceeds N_R, an opinion is normalised, in the general multinomial case, by $\frac{N_r}{\sum_{i=1}^{m} \alpha_i}$.
2. *Threshold parameter* t_S: A parameter that limits the impact of an opinion, based on the relative rank of the recommender reporting that opinion.

The Sybil resistant consensus operator is then, in its multinomial extension, given as

Definition 47 (Sybil Resistant Extended Consensus)**.** Let the trust of A in recommenders $R_1, \ldots, R_n$ be given in opinions $o^A_{R_1}, \ldots, o^A_{R_n}$. Let $\delta^A_{R_i}$ be the discounting factor assigned to recommender R_i by A based on opinion $o^A_{R_i}$, according to Definition 33, p. 126. Furthermore, let recommenders $R_1, \ldots, R_n$ provide opinions $o^{R_i}_{P_j}$ as recommendations on potential trustee P_j. For each $o^{R_i}_{P_j}$, let $0 \leqslant (\alpha^{R_i}_{P_j})_1 + \ldots + (\alpha^{R_i}_{P_j})_m \leqslant N_R$; if $\sum_{k=1}^{m} (\alpha^{R_i}_{P_j})_k > N_R$, the opinion is normalised accordingly, by computing:

$$\mathrm{norm}((\alpha_1, \ldots, \alpha_m)^\alpha) = \begin{cases} (\alpha_1, \ldots, \alpha_m)^\alpha & \text{if } \sum_{i=1}^{m} \alpha_i \leqslant N_R \\ (\frac{N_R}{\sum_{i=1}^{m} \alpha_i} \cdot \alpha_1, \ldots, \frac{N_R}{\sum_{i=1}^{m} \alpha_i} \cdot \alpha_m)^\alpha & \text{else} \end{cases}$$

Additionally, let the individual recommenders be ordered according to their trustworthiness estimates, so that i, the rank of R_i, is determined by $E(o^A_{R_i})$ and $R_i < R_k$ if $E(o^A_{R_i}) > E(o^A_{R_k})$. Let t_s denote the threshold for Sybil attacks. Then the Sybil Resistant Extended Consesus operation is defined as:

$$\begin{aligned} &\mathrm{consensus}_{t_s}(o^A_{R_1}, \ldots, o^A_{R_n}; o^{R_1}_{P_j}, \ldots, o^{R_n}_{P_j}) = [o^A_{R_1}, o^{R_1}_{P_j}] \hat{\oplus} \ldots \hat{\oplus} [o^A_{R_n}, o^{R_n}_{P_j}] \\ &= \Big(\textstyle\sum_{i=1}^{n} \min(\delta^A_{R_i} \cdot (\alpha_1)^{R_i}_{P_j}, (1 - t_s) \cdot (\delta^A_{R_i})^i \cdot \frac{N_R}{\sum_{k=1}^{m} (\alpha_k)^{R_i}_{P_j}} \cdot (\alpha_1)^{R_i}_{P_j}) \\ &\qquad\qquad , \ldots , \\ &\textstyle\sum_{i=1}^{n} \min(\delta^A_{R_i} \cdot (\alpha_m)^{R_i}_{P_j}, (1 - t_s) \cdot (\delta^A_{R_i})^i \cdot \frac{N_R}{\sum_{k=1}^{m} (\alpha_k)^{R_i}_{P_j}} \cdot (\alpha_m)^{R_i}_{P_j}) \Big) \end{aligned}$$

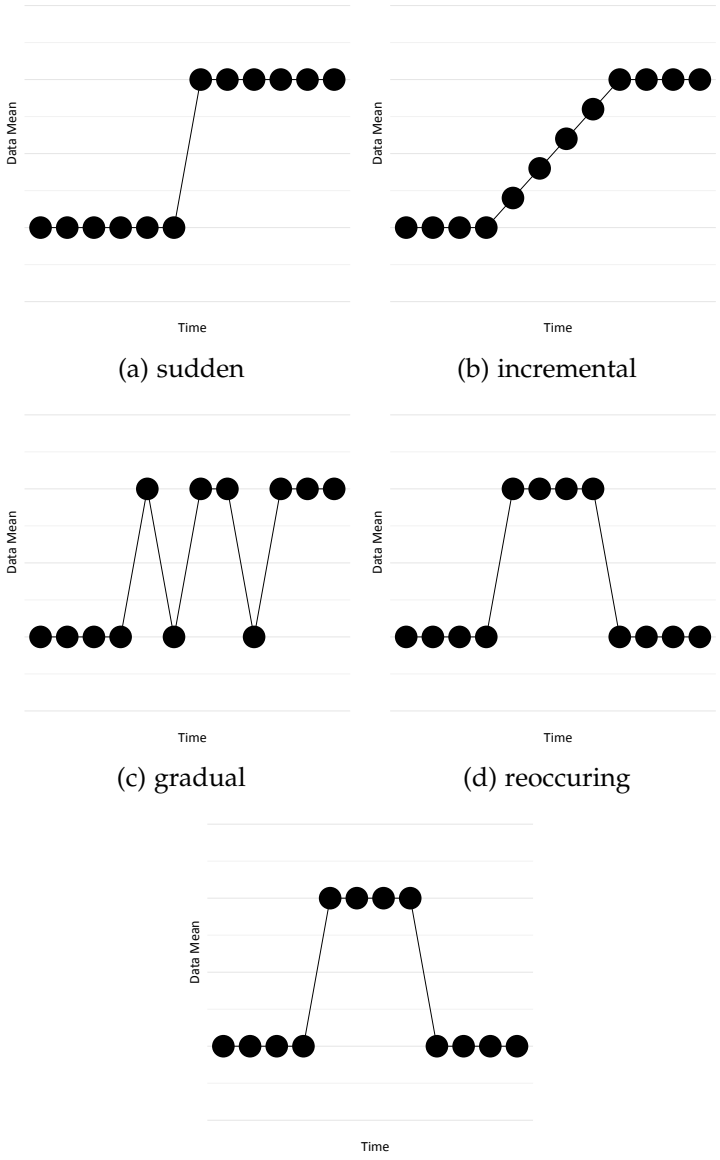
The parameters α_k represent the sufficient statistics *sum of observations in category* k, where $(\alpha_k)^{R_i}_{P_j}$ is the sum of observations in category k observed by entity R_i in interactions with entity P_j.

Sybil resistant consensus aggregation of truster A's opinion on trustee P_j, $o^A_{P_j}$, with recommendations $o^{R_1}_{P_j}, \ldots, o^{R_n}_{P_j}$ from recommenders $R_1, \ldots, R_n$ is thus achieved by computing the basic consensus (Definition 29, p. 113) of $o^A_{P_j}$ and the recommendations aggregated using the the Sybil resistant consensus operation:

$$\begin{aligned} &\mathit{consensus}\left(o^A_{P_j}, \mathit{consensus}_{t_s}(o^A_{R_1}, \ldots, o^A_{R_n}; o^{R_1}_{P_j}, \ldots, o^{R_n}_{P_j})\right) \\ &= o^A_{P_j} \oplus \left([o^A_{R_1}, o^{R_1}_{P_j}] \hat{\oplus} \ldots \hat{\oplus} [o^A_{R_n}, o^{R_n}_{P_j}]\right) \end{aligned}$$

E

FIGURES

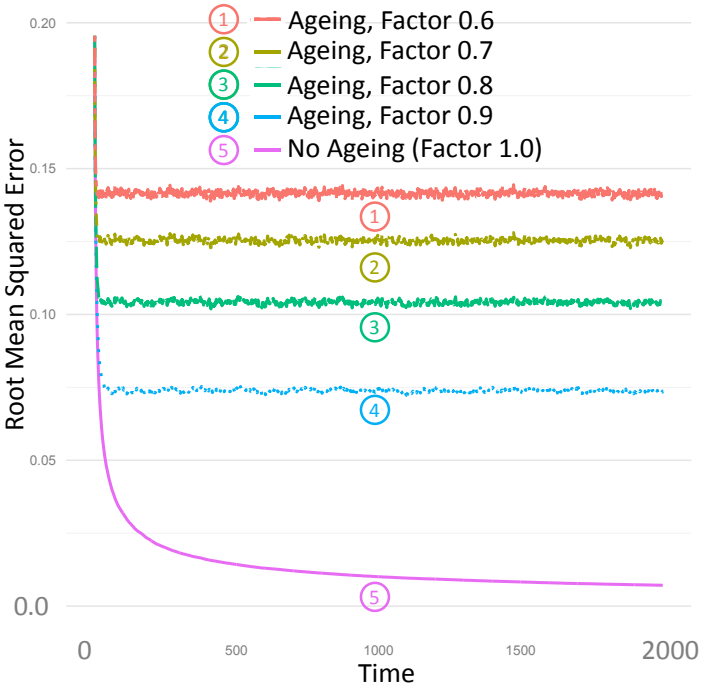


(a) sudden (b) incremental

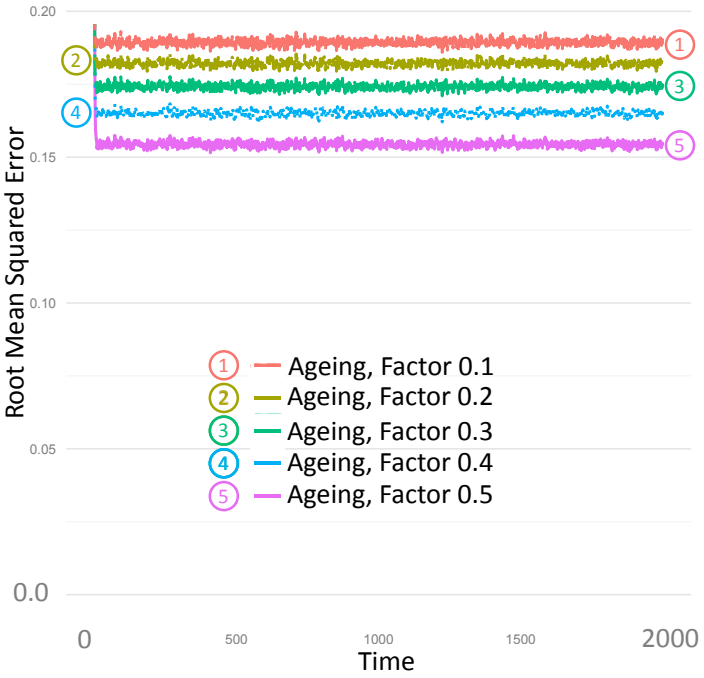
(c) gradual (d) reoccuring

(e) outlier

Figure 33: Patterns of changes in data over time (outlier not concept drift) [63]

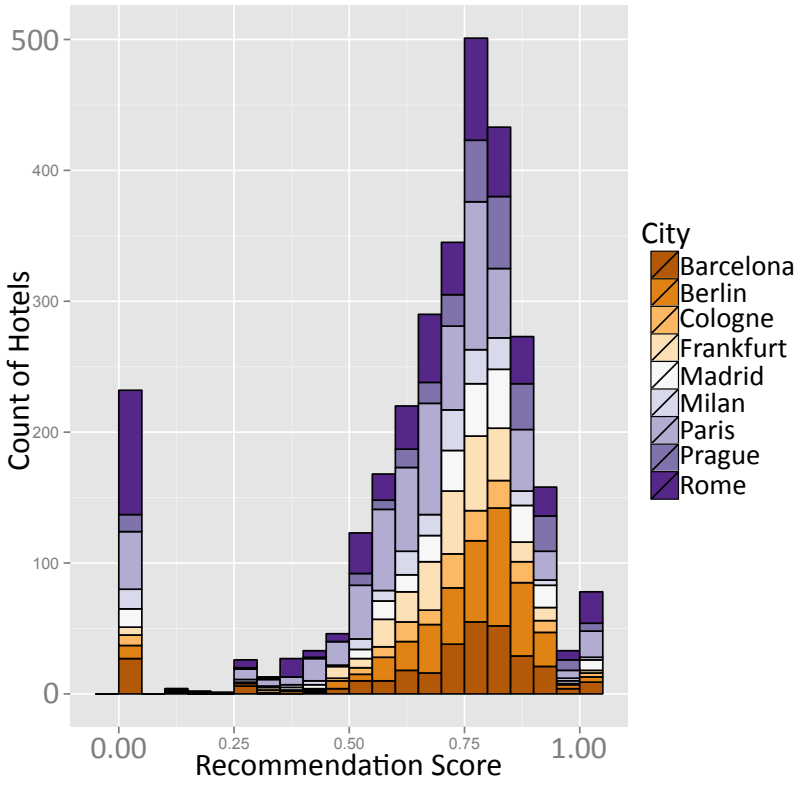


(a) Ageing, factors 0.6 to 1.0

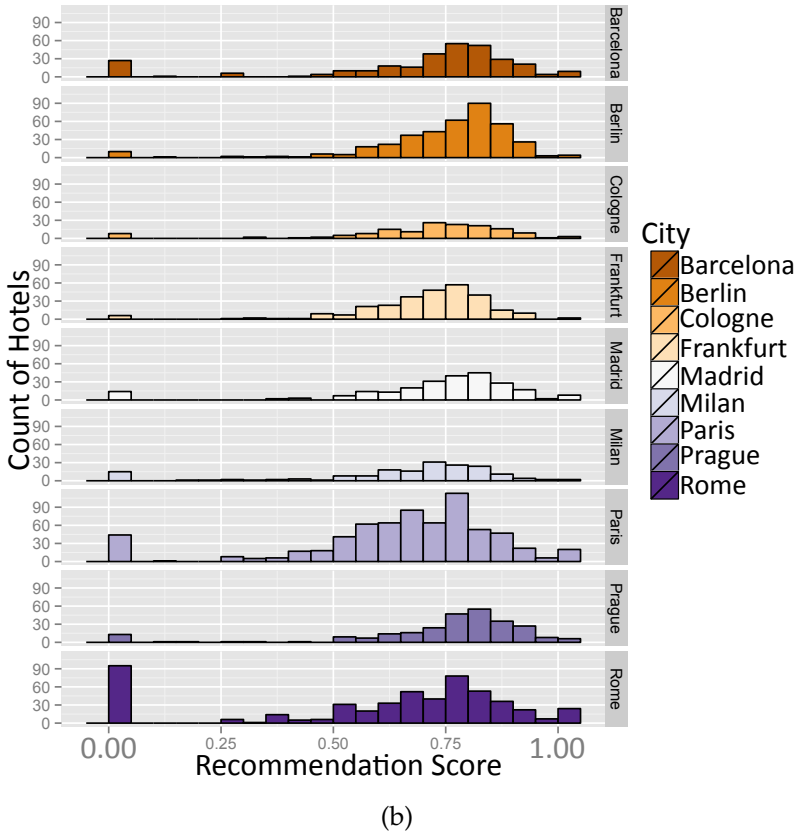


(b) Ageing, factors 0.1 to 0.5

Figure 34: Average accuracy, in terms of Root Mean Squared Error (Monte-Carlo simulation of a stationary Bernoulli Process, randomised p, $n = 2,000$, $10,000$ repeats).

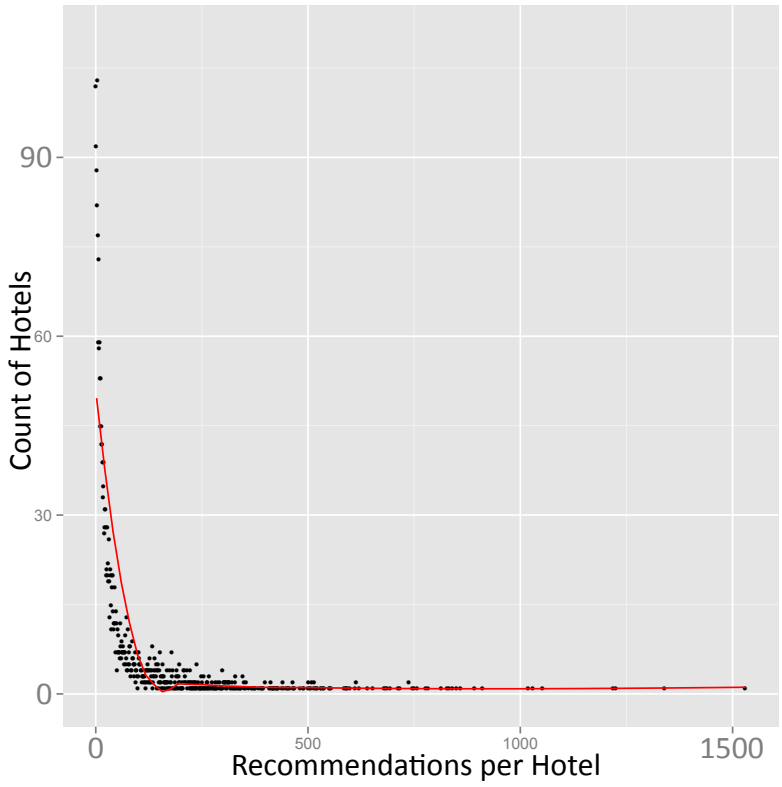


(a)

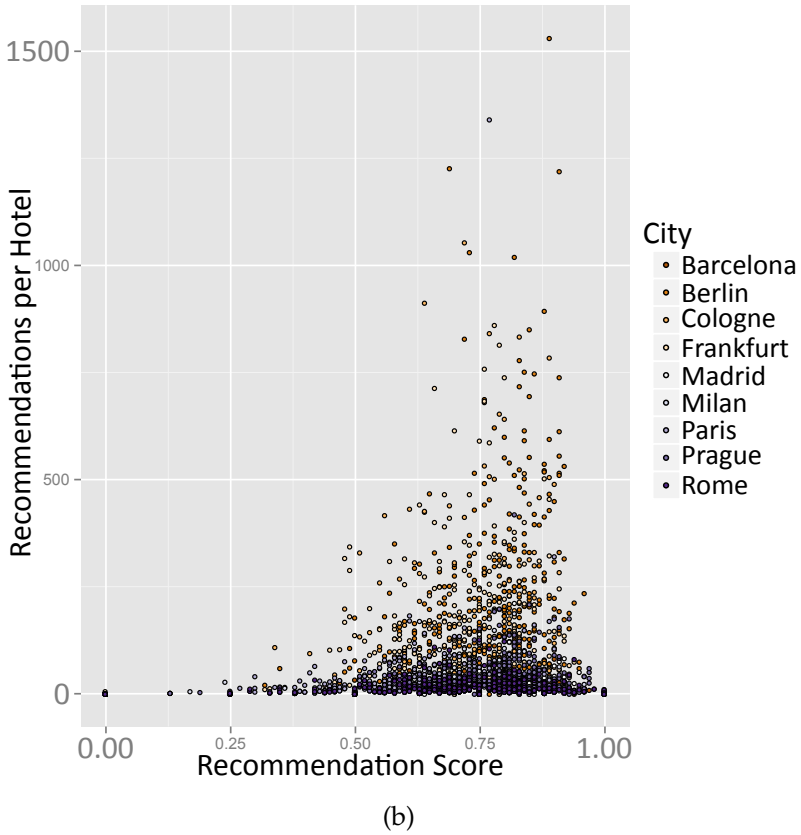


(b)

Figure 35: Aggregate recommendations in the hotel dataset.



(a)



(b)

Figure 36: Aggregate recommendations in the hotel dataset.

F

GOODNESS-OF-FIT MEASURES

Goodness-of-Fit measures according to [150] used for evaluation by using the HydroGOF *R*-package [208].

In the following, let $O = (o_1, o_2, \ldots, o_n)$ be a vector of observed values and $\hat{S} = (\hat{s}_1, \hat{s}_2, \ldots, \hat{s}_n)$ a vector of corresponding estimates. Let o_{max} be the largest, o_{min} the smallest element of O.

MEAN ERROR (ME)

$$ME = \frac{1}{n}\sum_{i=1}^{n}(\hat{s}_i - o_i)$$

MEAN ABSOLUTE ERROR (MAE)

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|\hat{s}_i - o_i|$$

MEAN SQUARED ERROR (MSE)

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(\hat{s}_i - o_i)^2$$

ROOT MEAN SQUARED ERROR (RMSE)

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{s}_i - o_i)^2}$$

NORMALISED ROOT MEAN SQUARED ERROR % (NRMSE%)

$$NRMSE\% = 100 \cdot \frac{\left(\sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{s}_i - o_i)^2}\right)}{nval}$$

with $nval$ a normalisation value:

$$nval = \begin{cases} sd(o_i) & \text{Standard Deviation of the observations (default)} \\ o_{max} - o_{min} & \end{cases}$$

PERCENT BIAS (PBIAS)

$$PBIAS = 100 \cdot \frac{\sum_{i=1}^{n}(\hat{s}_i - o_i)}{\sum_{i=1}^{n} o_i}$$

RATIO OF RMSE TO THE STANDARD DEVIATION OF THE OBSERVATIONS (RSR)

$$RSR = \frac{RMSE}{sd(o_i)}$$

RATIO OF STANDARD DEVIATIONS BETWEEN ESTIMATES AND OBSERVED VALUES (RSD)

$$rSD = \frac{sd(\hat{s}_i)}{sd(o_i)}$$

NASH-SUTCLIFFE EFFICIENCY (NSE)

$$NSE = 1 - \frac{\sum_{i=1}^{n}(\hat{s}_i - o_i)^2}{\sum_{i=1}^{n}(\hat{o}_i - \bar{o})^2}$$

where

$$\bar{o} = \frac{1}{n}\sum_{i=1}^{n} o_i$$

MODIFIED NASH-SUTCLIFFE EFFICIENCY (MNSE)

$$NSE = 1 - \frac{\sum_{i=1}^{n}|\hat{s}_i - o_i|^j}{\sum_{i=1}^{n}|\hat{o}_i - \bar{o}|^j}$$

where

$$\bar{o} = \frac{1}{n}\sum_{i=1}^{n} o_i$$

and $j = 1$ as the package default setting in HydroGOF.

INDEX OF AGREEMENT (D)

$$d = 1 - \frac{\sum_{i=1}^{n}(\hat{s}_i - o_i)^2}{\sum_{i=1}^{n}(|\hat{s}_i - \bar{o}| + |\hat{o}_i - \bar{o}|)^2}$$

where

$$\bar{o} = \frac{1}{n}\sum_{i=1}^{n} o_i$$

MODIFIED INDEX OF AGREEMENT (MD)

$$md = 1 - \frac{\sum_{i=1}^{n} (\hat{s}_i - o_i)^j}{\sum_{i=1}^{n} |\hat{s}_i - \bar{o}| + |\hat{o}_i - \bar{o}|^j}$$

where

$$\bar{o} = \frac{1}{n} \sum_{i=1}^{n} o_i$$

and $j = 1$ as the package default setting in HydroGOF.

BIBLIOGRAPHY

[1] Alfarez Abdul-Rahman and Stephen Hailes. Supporting Trust in Virtual Communities. In *Proceedings of the Hawaii's International Conference on Systems Sciences, Maui Hawaii*, 2000.

[2] Alan Agresti. A Survey of Exact Inference for Contingency Tables. *Statistical Science*, 7(1):131–153, 1992.

[3] Alan Agresti. *An Introduction to Categorical Data Analysis*. Wiley-Interscience, 2007.

[4] Alan Agresti and Brent A. Coull. Approximate is Better than "Exact" for Interval Estimation of Binomial Proportions. *The American Statistician*, 52:119–126, 1998.

[5] Alan Agresti, Cyrus R. Mehta, and Nitin R. Patel. Exact Inference for Contingency Tables with Ordered Categories. *Journal of the American Statistical Association*, 85(410):453–458, 1990.

[6] Dane Archer and Robin M. Akert. Words and Everything Else: Verbal and Nonverbal Cues in Social Interpretation. *Journal of Personality and Social Psychology*, 35(6):443–449, 1977.

[7] Kenneth Jospeh Arrow. *The Limits of Organization*. Norton, New York, 1974.

[8] Donovan Artz and Yolanda Gil. A survey of trust in computer science and the Semantic Web. *Web Semantics: Science, Services and Agents on the World Wide Web*, 5(2):58–71, June 2007. ISSN 15708268.

[9] Richard C. Atkinson, Gordon H. Bower, and Edward J. Crothers. *An Introduction to Mathematical Learning Theory*. Wiley, New York, 1965.

[10] Bernard Barber. *The Logic and Limits of Trust*. Rutger University Press, New Brunswick, 1983.

[11] Robert M. Bell, Yehuda Koren, and Chris Volinsky. All Togehter Now: A Perspective on the NETFLIX PRIZE. *Chance*, 23 (1):24–29, 2010.

[12] James O Berger. *Statistical Decision Theory and Bayesian Analysis*. Springer, 1985.

[13] José M. Bernardo and Adrian F. M. Smith. *Bayesian Theory*. Wiley & Sons, 1994.

[14] Ellen S. Berscheid. Interpersonal Relationships. *Annual Review of Psychology*, 45:79–129, 1994.

[15] Mark Best and Duncan Neuhauser. Walter A Shewhart, 1924, and the Hawthorne factory. *Quality and Safety in Health Care*, 15(2):142–143, 2006.

[16] Gérard Biau and Luc Devroye. On the Layered Nearest Neighbour Estimate, the Bagged Nearest Neighbor Estimate and the Random Forest Method in Regression and Classification. *Journal of Multivariate Analysis*, 101:2499–2518, 2010.

[17] Gérard Biau, Luc Devroye, and Gábor Lugosi. Consistency of Random Forests and Other Averaging Classifiers. *Journal of Machine Learning Research*, 9:2015–2033, 2008.

[18] Holger Billhardt, Ramón Hermoso, Sascha Ossowski, and Roberto Centeno. Trust-based service provider selection in open environments. In *Proceedings of the 2007 ACM Symposium on Applied Computing*, SAC '07, pages 1375–1380, New York, NY, USA, 2007. ACM. ISBN 1-59593-480-4.

[19] Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

[20] William M. Bolstad. *Introduction to Bayesian Statistics*. John Wiley & Sons, 2005.

[21] Susan D. Boon and John G. Holmes. The Dynamics of Interpersonal Trust: Resolving Uncertainty in the Face of Risk. In R. Hinde and J. Groebel, editors, *Cooperation and Prosocial Behavior*, pages 190–211. Cambridge University Press, Cambridge, 1991.

[22] Patrick D. Bourke. Sample size and the Binomial CUSUM Control Charts: the case of 100% inspection. *Metrika*, 53:51–70, 2001.

[23] W. John Braun. Run length distributions for estimated attributes charts. *Metrika*, 50(2):121–129, 1999.

[24] Leo Breiman. Bagging Predictors. *Machine Learning*, 24:123–140, 1996.

[25] Leo Breiman. Random Forests. *Machine Learning*, 45:5–32, 2001.

[26] Leo Breiman, Jerome H. Friedman, Charles J. Stone, and R. A. Olshen. Classification and Regression Trees. Technical report, Wadsworth & Brooks/Cole Advanced Books and Software, 1984.

[27] Lawrence D. Brown, T. Tony Cai, and Anirban DasGupta. Interval Estimation for a Binomial Proportion. *Statistical Science*, 16 (2):101–117, 2001. ISSN 08834237.

[28] Sonja Buchegger. *Coping with misbehavior in mobile ad-hoc networks*. Doctoral thesis, École Polytechnique Fédérale de Lausanne, 2004.

[29] Sonja Buchegger and Jean-Yves Le Boudec. A Robust Reputation System for P2P and Mobile Ad-hoc Networks. In *Proceedings of P2PEcon*, 2004.

[30] Judee K. Burgoon, Joseph A. Bonito, Artemio Ramirez, Norah E. Dunbar, Karadeen Kam, and Jenna Fischer. Testing the Interactivity Principle: Effects of Mediation, Propinquity, and Verbal and Nonverbal Modalities in Interpersonal Interaction. *Journal of Communication*, 52(3):657–677, 2002. ISSN 1460-2466. doi: 10.1111/j.1460-2466.2002.tb02567.x. URL http://dx.doi.org/10.1111/j.1460-2466.2002.tb02567.x.

[31] Chris Burnett, T.J. Norman, and Katia Sycara. Bootstrapping Trust Evaluations through Stereotypes. In *Proceedings. of 9th International Conference on Autonomous Agents and Multiagent Systems*, pages 241–248, 2010.

[32] Chris Burnett, T.J. Norman, and Katia Sycara. Sources of Stereotypical Trust in Multi-Agent Systems. In *Proceedings of the 14th International Workshop on Trust in Agent Societies*, page 25, 2011.

[33] Robert R. Bush and Frederick Mosteller. *Stochastic Models of Learning*. Wiley, New York, 1955.

[34] David G. Carnevale and Barton Wechsler. Trust in the Public Sector: Individual and Organizational Determinants. *Administration and Society*, 23:471–494, 1992.

[35] George Casella and Roger L. Berger. *Statistical Inference*. Duxbury Press, 2002.

[36] Cristiano Castelfranchi and Rino Falcone. Trust is much more than Subjective Probability: Mental Components and Sources of Trust. In *Proceedings of the 33rd Annual Hawaii International Conference on System Sciences*, 2000.

[37] Cynthia L. Corritore, Beverly Kracher, and Susan Wiedenbeck. On-line Trust: Concepts, Evolving Themes, a Model. *International Journal of Human-Computer Studies, Trust and Technology*, 58(6):737–758, 2003.

[38] Martin J. Crowder. Maximum Likelihood Estimation for Dependent Observations. *Journal of the Royal Statistical Society. Series B (Methodological)*, 38(1):45–53, 1976.

[39] Stephen C. Currall and Andrew C. Inkpen. On the Complexity of Organizational Trust: A Multi-Level Co-Evolutionary Perspective and Guidelines for Future Research. In R. Bachmann and A. Zaheer, editors, *Handbook of Trust Research*, pages 235–246. Edward Elgar Publishing, Northampton, Mass., 2006.

[40] Cassio P. De Campos and Alessio Benavoli. Inference with multinomial data: why to weaken the prior strength. In *Proceedings of the Twenty-Second international joint conference on Artificial Intelligence - Volume Volume Three*, IJCAI'11, pages 2107–2112. AAAI Press, 2011. ISBN 978-1-57735-515-1.

[41] Chrysanthos Dellarocas. Sanctioning Reputation Mechanisms in Online Trading Environments with Moral Hazard. MIT Sloan Working Paper 4297-03, MIT Sloan School of Management, July 2004.

[42] Zoran Despotovic and Karl Aberer. A Probabilistic Approach to Predict Peers' Performance in P2P Networks. In *In: 8th Intl Workshop on Cooperative Information Agents*, pages 62–76. Springer, 2004.

[43] Zoran Despotovic and Karl Aberer. P2P Reputation Management : Probabilistic Estimation vs Social Networks. *Computer Networks*, 50:485–500, 2006.

[44] Morton Deutsch. *The Resolution of Conflict: Constructive and Destructive Processes*. Yale University Press, New Haven, Conn., 1973.

[45] Thorsten Dikmann. Entwicklung eines Frameworks zur Simulation von Kontaktausbreitung in komplexen Netzwerken. Master's thesis, Westfälische Wilhelms-Universität Münster, 2010.

[46] Roger Dingledine, Michael J Freedman, and David Molnar. *Peer-to-Peer: Harnessing the Power of Disruptive Technologies*, chapter Accountability Measures for Peer-to-Peer Systems, pages 271–340. O'Reilly and Associates, 2001.

[47] Norman R. Draper and Harry Smith. *Applied Regression Analysis*. Wiley Series in Probability and Statistics. Wiley, 3rd edition edition, 1998.

[48] Richard O. Duda, Peter E. Hart, and David G. Stork. *Pattern Classification*. Wiley, 2001.

[49] Edward J. Dudewicz and Satya Mishra. *Modern Mathematical Statistics*. John Wiley & Sons, Inc., New York, NY, USA, 1988. ISBN 0-47-181472-5.

[50] Benjamin Edelman. Adverse selection in online trust certifications. In *Proceedings of the 11th International Conference on Electronic Commerce*, pages 205–212, 2009. ISBN 9781605585864.

[51] Dag Elgesem. Normative Structures in Trust Management. In K. Stølen, W. H. Winsborough, F. Martinelli, and F. Massacci, editors, *iTrust 2006*, pages 48–61. Springer, Heidelberg, 2006.

[52] Hui Fang, Jie Zhang, Murat Sensoy, and Nadia M. Thalmann. A Generalized Stereotypical Trust Model. In *IEEE TrustCom-11*, pages 698–705. IEEE, 2012.

[53] Tom Fawcett. An introduction to ROC analysis. *Pattern Recognition Letters*, 27:861–874, 2006.

[54] Michal Feldman, Christos Papadimitriou, John Chuang, and Ion Stoica. Free-Riding and Whitewashing in Peer-to-Peer Systems. In *Proceedings of the ACM SIGCOMM Workshop on Practice and Theory of Incentives in Networked Systems (PINS '04)*, pages 228–236, 2004.

[55] Ronald A. Fisher. On the interpretation of chi-squared from contingency tables, and the calculation of P. *Journal of the Royal Statistical Society* , 85(1):87–94, 1922.

[56] Ronald A. Fisher. Two New Properties of Mathematical Likelihood. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 144(852):285–307, 1934.

[57] Alan Fox. *Beyond Contract: Work, Power and Trust Relations*. Faber, London, 1974.

[58] G. H. Freeman and T. R. Halton. Note on exact treatment of contingency, goodness-of-fit and other problems of significance. *Biometrika*, 38:141–149, 1951.

[59] Eric J. Friedman and Paul Resnick. The Social Cost of Cheap Pseudonyms. *Journal of Economics & Management Strategy*, 10(2) (2):173–199, June 2001. ISSN 1058-6407.

[60] Bela A. Frigyik, Amol Kapila, and Maya R. Gupta. Introduction to the Dirichlet Distribution and Related Processes. Technical Report UWEETR-2010-0006, University of Washington, 2010.

[61] Francis Fukuyama. *Trust: The Social Virtues and the Creation of Prosperity*. Free Press, New York, 1996.

[62] John J. Gabarro. The Development of Trust, Influence and Expectations. In A. G. Athos and J. J. Gabarro, editors, *Interpersonal Behavior: Communication and Understanding in Relationships*, pages 290–303. Prentice-Hall, Englewood Cliffs, 1978.

[63] João Gama, Indrė Žliobaitė, Albert Bifet, Mykola Pechenizkiy, and Abdelhamid Bouchachia. A Survey on Concept Drift. *ACM Computing Surveys*, 1:1–35, 2013.

[64] Diego Gambetta. Can We Trust Trust? In D. Gambetta, editor, *Trust: Making and Breaking Cooperative Relations*, pages 213–237. Basil Blackwell, Oxford, 1988.

[65] Saurabh Ganeriwal, Laura K. Balzano, and Mani B. Srivastava. Reputation-based framework for high integrity sensor networks. *ACM Transactions on Sensor Networks*, 4(3):1–37, 2008. ISSN 15504859.

[66] Subhashis Ghosal. *Bayesian Nonparametrics*, chapter The Dirichlet Process, Related Priors and Posterior Asymptotics, pages 36–38. Cambridge University Press, 2010.

[67] Geof H. Givens and Jennifer A. Hoeting. *Computational Statistics*. Wiley Series in Computational Statistics. Wiley, 2nd edition edition, 2012.

[68] Jennifer A. Golbeck. *Computing and Applying Trust in Web-based Social Networks*. Doctoral thesis, University of Maryland, 2005. URL http://129.2.17.93/drum/handle/1903/2384.

[69] Robert T. Golembiewski and Mark McConkie. The Centrality of Interpersonal Trust in Group Processes. In G. L. Cooper, editor, *Theories of Group Processes*, pages 131–185. John Wiley & Sons, London, 1975.

[70] Leo A. Goodman. On Simultaneous Confidence Intervals for Multinomial Proportions. *Technometrics*, 7:247–254, 1965.

[71] Ronald L. Graham, Donald E. Knuth, and Oren Patashnik. *Concrete Mathematics: A Foundation for Computer Science*. Addison-Wesley, Reading, MA, 1990.

[72] Jon E. Grahe and Frank J. Bernieri. The Importance of Nonverbal Cues in Judging Rapport. *Journal of Nonverbal Behaviour*, 23(4):253–269, 1999.

[73] Tyrone Grandison and Morris Sloman. A Survey of Trust in Internet Applications. *IEEE Communications and Survey*, 3(4):2–16, 2000.

[74] Clive W.J. Granger. Some Properties of Time Series Data and Their Use in Econometric Model Specification. *Journal of Econometrics*, 16:121–130, 1981.

[75] Sheikh Mahbub Habib. *Trust Establishment Mechanisms for Distributed Service Environments*. PhD thesis, Technische Universität Darmstadt, 2013.

[76] Sheikh Mahbub Habib, Sascha Hauke, Sebastian Ries, and Max Mühlhäuser. Trust as a Facilitator in Cloud Computing: A Survey. *Journal of Cloud Computing: Advances, Systems and Applications*, 1:1–19, 2012.

[77] Sheikh Mahbub Habib, Sebastian Ries, Sascha Hauke, and Max Mühlhäuser. Fusion of Opinions under Uncertainty and Conflict–Trust Assessment for Cloud Marketplaces. In *Proceedings of IEEE TrustCom-12*, 2012.

[78] Sheikh Mahbub Habib, Sebastian Ries, Max Mühlhäuser, and Prabhu Varikkattu. Towards a Trust Management System for Cloud Computing Marketplaces: using CAIQ as a trust information source. *Security and Communication Networks Journal*, 2013.

[79] Chung-Wei Hang and Munindar P. Singh. Trustworthy Service Selection and Composition. *ACM Transactions on Autonomous and Adaptive Systems (TAAS)*, 6(1):5, 2011. ISSN 1556-4665.

[80] Sascha Hauke, Martin Pyka, Markus Borschbach, and Dominik Heider. Harnessing Recommendations from Weakly Linked Neighbors in Reputation-based Trust Formation. In *Proceedngs of the 2010 International Conference on Cyberworlds*, 2010.

[81] Sascha Hauke, Martin Pyka, Markus Borschbach, and Dominik Heider. Reputation-based Trust Diffusion in Complex Socio-Economic Networks. In *Information Retrieval and Mining in Distributed Environments*, chapter Reputation. Springer, Berlin-Heidelberg, 2010.

[82] Sascha Hauke, Martin Pyka, Markus Borschbach, and Dominik Heider. Augmenting Reputation-Based Trust Metrics with Rumor-Like Dissemination of Reputation Information. *IFIP Advances in Information and Communication Technology*, 330:136–147, 2010.

[83] Sascha Hauke, Martin Pyka, and Dominik Heider. Group-Agreement as a Reliability Measure for Witness Recommendations in Reputation-based Trust Protocols. *Transactions on Computational Science, Lecture Notes on Computational Science*, XII:231–255, 2011.

[84] Sascha Hauke, Florian Volk, Sheikh Mahbub Habib, and Max Mühlhäuser. Integrating Indicators of Trustworthiness into Reputation-based Trust Models. In *Proceedings of the 6th IFIP WG 11.11 International Conference, IFIPTM 2012*, 2012.

[85] Sascha Hauke, Sebastian Biedermann, Max Mühlhäuser, and Dominik Heider. On the Application of Supervised Machine

Learning to Trustworthiness Assessment. In *Proceedings of the 12th IEEE International Conference on Trust, Security and Privacy in Computing and Communications (IEEE TrustCom-13)*, 2013.

[86] Sascha Hauke, Sebastian Biedermann, Max Mühlhäuser, and Dominik Heider. On the Application of Supervised Machine Learning to Trustworthiness Assessment. Technical Report TUD-CS-2013-0050, TR-014, Technische Universität Darmstadt, 2013.

[87] Douglas M. Hawkins, Peihua Qiu, and CW Kang. The Changepoint Model for Statistical Process Control. *Journal of Quality Technology*, 35(4):355–366, 2003.

[88] Risto D.H. Heijmans and Jan R. Magnus. Consistent Maximum-Likelihood Estimation With Dependent Observations – The General (Non-Normal) Case and the Normal Case. *Journal of Econometrics*, 32:253–285, 1986.

[89] Joseph M. Hilbe. *Logistic Regression Models*. Chapman and Hall, CRC Press, 2009.

[90] Yosef Hochberg and Tamhane Ajit C. *Multiple Comparison Procedures*. Wiley, New York, 1987.

[91] Nan Hu, Paul A. Pavlou, and Jie Zhang. Overcoming the J-shaped Distribution of Product Reviews. *Communications of the ACM*, 52 (10):144–147, 2009.

[92] Gonzalo Huerta-Canepa, Han Seungwook, Dongman Lee, and Byoungoh Kim. A Place-aware Stereotypical Trust Supporting Scheme. In *Proceedings of the 12th IEEE International Conference on Trust, Security and Privacy in Computing and Communications*, pages 821–828, 2013.

[93] Trung Dong Huynh. *Trust and Reputation in Open Multi-Agent Systems*. PhD thesis, University of Southampton, 2006.

[94] Trung Dong Huynh, Nicholas R. Jennings, and Nigel R. Shadbolt. An integrated trust and reputation model for open multi-agent systems. *Autonomous Agents and Multi-Agent Systems*, 13 (2):119–154, 2006.

[95] FBI IC3. 2013 Internet Crime Report. Technical report, Federal Bureau of Investigation, Internet Crime Complaint Center, 2013.

[96] Edwin T. Jaynes. Prior Probabilities. *IEEE Transactions on Systems Science and Cybernetics*, SSC-4:227–241, 1968.

[97] Edwin T. Jaynes. Confidence Intervals vs Bayesian Intervals. *Foundations of Probability Theory, Statistical Inference, and Statistical Theories of Science*, II:175–257, 1976.

[98] Edwin T. Jaynes. *Papers on Probability, Statistics, and Statistical Physics*. Reidel, Dordrecht, 1983.

[99] Harold Jeffreys. *Theory of Probability*. Oxford University Press, London, 3rd edition, 1961.

[100] Emily M. Jin, Michelle Girvan, and M. E. J. Newman. Structure of Growing Social Networks. *Physical Review E*, 64(4):46132, September 2001. doi: 10.1103/PhysRevE.64.046132.

[101] Norman L. Johnson, Samuel Kotz, and N. Balakrishnan. *Continuous Univariate Distributions*, volume 1. John Wiley & Sons, 2nd edition, 1994.

[102] Audun Jøsang. A Subjective Metric of Authentication. In *Proceedings of the 5th European Symposium on Research in Computer Security (ESORICS)*, volume 1485 of *Lecture Notes in Computer Science (LNCS)*, pages 329–344. Springer-Verlag, Heidelberg, 1998.

[103] Audun Jøsang. A Logic for Uncertain Probabilities. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 9 (3):279–311, 2001.

[104] Audun Jøsang. Probabilistic Logic Under Uncertainty. In *Proceedings of Computing: The Australian Theory Symposium (CATS'07)*, 2007.

[105] Audun Jøsang. Fission of Opinions in Subjective Logic. In *Proceedings of the 12th International Conference on Information Fusion (FUSION 2009)*, 2009.

[106] Audun Jøsang. Subjective Logic. Technical report, University of Oslo, 2013. URL http://folk.uio.no/josang/papers/subjective_logic.pdf.

[107] Audun Jøsang and Jochen Haller. Dirichlet Reputation Systems. In *Proceedings of the 2007 International Conference on Availability, Reliability and Security (ARES)*, pages 112–119. IEEE Computer Society, 2007.

[108] Audun Jøsang and Roslan Ismail. The Beta Reputation System. In *Proceedings of the 15th Bled Electronic Commerce Conference*, 2002.

[109] Audun Jøsang and S. Lo Presti. Analysing the Relationship between Risk and Trust. In *Proceedings of the Second International Conference on Trust Management (iTrust 2004)*, pages 135–145. Springer, 2004.

[110] Audun Jøsang and David McAnally. Multiplication and Comultiplication of Beliefs. *International Journal of Approximate Reasoning*, 38:19–55, 2005.

[111] Audun Jøsang, Claudia Keser, and Theo Dimitrakos. Can We Manage Trust? In *Proceedings of the 3rd International Conference on Trust Management*, pages 93–107, Berlin, 2005. Springer.

[112] Audun Jøsang, Claudia Keser, and Theo Dimitrakos. Can we manage trust? In *Proceedings of the Third International Conference on Trust Management (iTrust), Versailes*, pages 93–107, Berlin, 2005. Springer-Verlag.

[113] Audun Jøsang, Roslan Ismail, and Colin Boyd. A Survey of Trust and Reputation Systems for Online Service Provision. *Decision Support Systems*, 43 (2):618–644, 2007.

[114] Audun Jøsang, Touhid Bhuiyan, Yue Xu, and Clive Cox. Combining Trust and Reputation Management for Web-Based Services. In *Proceedings of the 5th international conference on Trust, Privacy and Security in Digital Business (TrustBus '08)*, 2008.

[115] Sepandar D. Kamvar, Mario T. Schlosser, and Hector Garcia-Molina. The Eigentrust algorithm for reputation management in P2P networks. In *Proceedings of the Twelfth International Conference on World Wide Web (WWW '03)*, 2003. ISBN 1581136803.

[116] Robert E Kass and Adrian E Raftery. Bayes Factors. *Journal of the American Statistical Association*, 90 (430):773–795, 1995.

[117] Robert E. Kass and E. Wasserman. The Selection of Prior Distributions by Formal Rules. *Journal of the American Statisitical Association*, 91 (435):1343–1366, 1996.

[118] Michael Kinateder. *Fundamental Models and Algorithms for a Distributed Reputation System*. PhD thesis, Universität Stuttgart, 2008.

[119] Michael Kinateder and Kurt Rothermel. Architecture and Algorithms for a Distributed Reputation System. *Proceedings of the First International Conference on Trust Management (iTrust 2003)*, LNCS 2692:1–16, 2003.

[120] Michael Kinateder, Ernesto Baschny, and Kurt Rothermel. Towards a Generic Trust Model. *Proceedings of the Third International Conference on Trust Management (iTrust 2005)*, LNCS 3477: 177–192, 2005.

[121] Karl Krukow, Mogens Nielsen, and Vladimiro Sassone. Trust Models in Ubiquitous Computing. *Philosophical Transactions. Series A, Mathematical, Physical, and Engineering Sciences*, 366(1881): 3781–93, October 2008. ISSN 1364-503X.

[122] Jochen Kruppa, Yufeng Liu, Gérard Biau, Michael Kohler, Inke R. König1, James D. Malley, and Andreas Ziegler. Probability estimation with machine learning methods for dichotomous and multicategory outcome: Theory. *Biometrical Journal*, 4:534–563, 2014.

[123] Terran Lane and Carla E. Brodley. Approaches to Online Learning and Concept Drift for User Identification in Computer Security. In *Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining, KDD*, pages 259–263. AAAI Press, 1998.

[124] Pierre-Simon Laplace. *Essai philosophique sur les probabilités*. 1820.

[125] Helen Leggatt. BizReport : Law & Regulation : April 01, 2009 – After two years in decline, Internet fraud complaints on the rise. online, April 2009. URL http://www.bizreport.com/2009/04/after_two_years_in_decline_internet_fraud_complaints_on_the.html.

[126] E.L. Lehmann and George Casella. *Theory of Point Estimation*, volume XXVI of *Springer Texts in Statistics*. Springer, 2nd edition, 1998.

[127] J. David Lewis and Andrew J. Weigert. Trust as a Social Reality. *Social Forces*, 63:967–985, 1985.

[128] Xin Liu, Anwitaman Datta, Krzysztof Rzadca, and Ee-Peng Lim. Stereotrust: A Group Based Personalized Trust Model. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management*, pages 7–16. ACM, 2009.

[129] Xin Liu, Gilles Trédan, and Anwitaman Datta. A Generic Trust Framework for Large-Scale Open Systems Using Machine Learning. *CoRR*, abs/1103.0086, 2011.

[130] Sharon L. Lohr. *Sampling: Design and Analysis*. Duxbury Press, Pacific Grove, 1999.

[131] Robert Duncan Luce. *Utility of gains and losses: Measurement-theoretical and experimental approaches*. Lawrence Erlbaum Associates Publishers, Mahwah, NJ, 2000.

[132] Niklas Luhmann. Familiarity, Confidence and Trust: Problems and Alternatives. In Diego Gambetta, editor, *Trust: Making and Breaking of Cooperative Relations*, pages 94–107. Basil Blackwell, Oxford, 1988.

[133] James D. Malley, Jochen Kruppa, Abhijit Dasgupta, Karen G. Malley, and Andreas Ziegler. Probability Machines: Consistent Probability Estimation Using Nonparametric Learning Machines. *Methods of Information in Medicine*, 51(1):74–81, 2012.

[134] Kerry L. Marsh, Dawn M. Hart-O'Rourke, and Deana L. Julka. The Persuasive Effects of Verbal and Nonverbal Information in a Context of Value Relevance. *Personallity and Social Psychology Bulletin*, 23(6):563–579, 1997.

[135] Stephen Marsh. *Formalising Trust as a Computational Concept*. PhD thesis, Department of Computing Science and Mathematics, University of Stirling, 1994.

[136] Stephen Marsh and Mark R. Dibben. Trust, Untrust, Distrust and Mistrust – An Exploration of the Dark(er) Side. In *Proceedings of Third iTrust International Conference (iTrust 2005)*, pages 17–33, 2005.

[137] Warren L. May and William D. Johnson. Constructing two-sided simultaneous confidence intervals for multinomial proportions for small counts in a large number of cells. *Journal of Statistical Software*, 5(6):1–24, 5 2000. ISSN 1548-7660.

[138] Roger C. Mayer, James H. Davis, and F. David Schoorman. An Integrative Model of Organizational Trust. *Academy of Management Review*, 20(3):709–734, 1995.

[139] Dina Mayzlin, Yaniv Dover, and Judith Chevalier. Who Gave That Hotel Five Stars? The Concierge... *Harvard Business Review*, September, 2012.

[140] Dina Mayzlin, Yaniv Dover, and Judith Chevalier. Promotional Reviews: An Empirical Investigation of Online Review Manipulation. Working Paper 18340, National Bureau of Economic Research, Cambridge, MA, August 2012.

[141] Katelyn Y. A. McKenna and John A. Bargh. Plan 9 From Cyberspace:The Implications of the Internet for Personality and Social Psychology. *Personality and Social Psychology Review*, 4(1): 57–75, 2000.

[142] D. Harrison McKnight and Norman L. Chervany. The Meanings of Trust. Technical report, University of Minnesota Management Information Systems Research Center, 1996.

[143] D. Harrison McKnight and Norman L. Chervany. Trust and Distrust Definitions: One Bite at a Time. In Cristiano Castelfranchi and Rino Falcone, editors, *Trust in Cyber-Societues*, pages 27–54. Springer, Berlin, 2001.

[144] D. Harrison McKnight, Larry L. Cummings, and Norman L. Chervany. Initial Trust Formation in New Organizational Relationships. *Academy of Management Review*, 23(3):473–490, 1998.

[145] D. Harrison McKnight, Vivek Choudhury, and Charles Kacmar. Developing and Validating Trust Measures for e-Commerce: An Integrative Typology. *Information Systems Research*, 13 (3)(3):334–359, 2002.

[146] Aneil K. Mishra. Organizational Responses to Crisis: The Centrality of Trust. In R. M. Kramer and T. R. Tyler, editors, *Trust in Organizations: Frontiers of Theory and Research*, pages 261–287. Sage, Thousand Oaks, 1996.

[147] Aneil K. Mishra and Karen E. Mishra. *Trust is Everything*. 2009.

[148] Alexander McFarlane Mood, Franklin A. Graybill, and Duane C. Boes. *Introduction to the Theory of Statistics*. McGraw-Hill, 3rd edition, 1973.

[149] Robert C. Moore. On Log-Likelihood-Ratios and the Significance of Rare Events. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 333–340, 2004.

[150] Daniel N. Moriasi, Jeffrey G. Arnold, Michael W. Van Liew, Ronald L. Bingner, R. Daren Harmel, and Tamie L. Veith. Model evaluation guidelines for systematic quantification of accuracy in watershed simulations. *Transactions of the American Society of Agricultural and Biological Engineers*, 50(3):885–900, 2007.

[151] Lik Mui. *Computational Models of Trust and Reputation: Agents, Evolutionary Games and Social Networks*. Doctoral dissertation, Massachusetts Institute of Technology, 2003.

[152] Lik Mui, Mojdeh Mohtashemi, and Ari Halberstadt. A computational model of trust and reputation. In *Proceedings of the 35th Hawaii International Conference on System Science*, pages 280–287, 2002.

[153] Tim Muller and Patrick Schweitzer. On Beta Models with Trust Chains. *Trust Management VII (IFIP Advances in Information and Communication Technology)*, 401:49–65, 2013.

[154] Kevin R. Murphy and Charles O. Davidshofer. *Psychological Testing : Principles and Applications*. Prentice Hall, 2005.

[155] Guy P. Nason. *Statistics in Volcanology*, chapter Stationary and Non-Stationary Time Series, pages 129–142. The Geological Society Publishing House, Bath, UK, 2006.

[156] Peter R. Nelson. Multiple Comparisons of Means Using Simultaneous Confidence Intervals. *Journal of Quality Technology*, 21: 232–241, 1989.

[157] Jerzy Neyman. Outline of a Theory of Statistical Estimation Based on the Classical Theory of Probability. *Philosophical Transactions of the Royal Society of London A*, 236:333–380, 1937.

[158] Bart Nooteboom. Trust as a Governance Device. In M. C. Casson and A. Godley, editors, *Cultural Factors in Economic Growth*. Springer, Heidelberg, 1999.

[159] John S. Oakland. *Statistical Process Control*. Butterworth-Heinemann, Oxford, 2003.

[160] P. S. Pawar, Muttukrishnan Rajarajan, Srijith Krishnan Nair, and Andrea Zisman. Trust Model for Optimized Cloud Services. *IFIP Advances in Information and Communication Technology*, 374: 97–112, 2012.

[161] Karl Pearson. On the Criterion that a given System of Deviations from the Probable in the Case of a Correlated System of Variables is such that it can be reasonably supposed to have arisen from Random Sampling. *Philosophical Magazine Series 5*, 50 (302):157–175, 1900.

[162] Rufus Pichler. Trust and Reliance - Enforcement and Compliance: Enhancing Consumer Confidence in the Electronic Marketplace. Juridical sciences master, Stanford University, 2000. URL http://www.oecd.org/dataoecd/0/18/1879122.pdf.

[163] Aleksey S. Polunchenko and Alexander G. Tartakovsky. State-of-the-Art in Sequential Change-Point Detection. *Methodology and Computing in Applied Probability*, 14 (3):649–684, 2011.

[164] Robert D. Putnam. *Making Democracy Work. Civil Traditions in Modern Italy*. Princeton University Press, Princeton, 1993.

[165] Charles P. Quesenberry and D. C. Hurst. Large Sample Simultaneous Confidence Intervals for Multinomial Proportions. *Technometrics*, 6:191–195, 1964.

[166] John R. Quinlan. Learning with Continuous Classes. In *Proceedings AI*, pages 343–348, 1992.

[167] Howard Raiffa and Robert Schlaifer. *Applied Statistical Decision Theory*. Division of Research, Graduate School of Business Administration, Harvard University, 1961.

[168] Sarvapali Ramchurn, Charles Sierra, Luis Godo, and Nicholas R. Jennings. Devising a Trust Model for Multi-Agent

Interactions Using Confidence and Reputation. *International Journal of Applied Artificial Intelligence*, 18:9–10, 2004.

[169] Calyampudi R. Rao. Large Sample Tests of Statistical Hypotheses Concerning Several Parameters with Application to Problems of Estimation. *Proceedings of the Cambridge Philosophical Society*, 44:50–57, 1948.

[170] Marion R. Reynolds and Zachary G. Stoumbos. A CUSUM Chart for Monitoring a Proportion When Inspecting Continuously. *Journal of Quality Technology*, 31:87–108, 1999.

[171] Elaine Rich. User Modeling via Stereotypes. *Cognitive Science*, 3 (4):329–354, 1979.

[172] Sebastian Ries. CertainTrust: a Trust Model for Users and Agents. In *Proceedings of the 2007 ACM Symposium on Applied Computing (SAC '07)*, 2007. ISBN 1-59593-480-4.

[173] Sebastian Ries. *Trust in Ubiquitous Computing*. Doctoral thesis, TU Darmstadt, 2009.

[174] Sebastian Ries. Extending Bayesian Trust Models Regarding Context-Dependence and User Friendly Representation. In *Proceedings of the 2009 ACM Symposium on Applied Computing*, pages 1294–1301, 2009.

[175] Sebastian Ries, Sheikh Mahbub Habib, Max Mühlhäuser, and Vijay Varadharajan. CertainLogic: A Logic for Modeling Trust and Uncertainty. In *Proceedings of the 4th International Conference on Trust and Trustworthy Computing (TRUST 2011)*. Springer, Jun 2011.

[176] W. H. Riker. The Nature of Trust. In J. T. Tedeschi, editor, *Perspectives on Social Power*, pages 63–81. Aldine, Chicago, 1971.

[177] Gordon J. Ross. *R-Package CPM: Sequential Parametric and Nonparametric Change Detection*, 01 2012.

[178] Gordon J. Ross, Dimitris K Tasoulis, and Niall M. Adams. Sequential monitoring of a Bernoulli sequence when the prechange parameter is unknown. *Computational Statistics*, 28(2): 463–479, 2012.

[179] Jordi Sabater. *Trust and Reputation for Agent Societies*. PhD thesis, Universitat Autonoma de Barcelona, 2003.

[180] Vladimiro Sassone, Karl Krukow, and Mogens Nielsen. Towards a Formal Framework for Computational Trust. *Formal Methods for Components*, pages 175–184, 2006.

[181] Vladimiro Sassone, Karl Krukow, and Mogens Nielsen. Towards a Formal Framework for Computational Trust. In FrankS. Boer, MarcelloM. Bonsangue, Susanne Graf, and Willem-Paul Roever, editors, *Formal Methods for Components and Objects*, volume 4709 of *Lecture Notes in Computer Science*, pages 175–184. Springer, 2007. ISBN 978-3-540-74791-8.

[182] George A. F. Seber and Alan J. Lee. *Linear Regression Analysis*. Wiley Series in Probability and Statistics. Wiley, 2nd edition edition, 2003.

[183] Glenn Shafer. *A Mathematical Theory of Evidence*. Princeton University Press, 1976.

[184] Walter Andrew Shewhart. *Economic Control of Quality of Manufactured Product*. D. Van Nostrand Company, Inc., 1931.

[185] Avinash Srinivasan, Joshua Teitelbaum, and Huigang Liang. *Algorithms and Protocols for Wireless Ad-Hoc and Sensor Networks*, chapter Reputation and Trust-based Systems for Ad Hoc and Sensor Networks, pages 375–402. Wiley & Sons, 2009.

[186] Stephen Mack Stigler. Thomas Bayes's Bayesian Inference. *Journal of the Royal Statistical Society*, Ser. A, 145:250–258, 1982.

[187] Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, Cambridge, MA, 1998.

[188] Piotr Sztompka. *Trust: A Sociological Theory*. Cambridge University Press, Cambridge, 1999.

[189] W. T. Luke Teacy, Jigar Patel, Nicholas R. Jennings, and Michael Luck. TRAVOS: Trust and Reputation in the Context of Inaccurate Information Sources. *Autonomous Agents and Multi-Agent Systems*, 12(2):183–198, 2006. ISSN 1387-2532.

[190] W. T. Luke Teacy, Michael Luck, Alex Rogers, and Nicholas R. Jennings. An Efficient and Versatile Approach to Trust and Reputation using Hierarchical Bayesian Modelling. *Artificial Intelligence*, 193:149–185, 2012.

[191] David J. Thomson. Jackknifing Multiple-Window Spectra. In *Proceedings of the Sixth IEEE International Conference on Acoustics, Speech and Signal Processing*, 1994.

[192] Elisabeth Topalidou and Stelios Psarakis. Review of Multinomial and Multiattribute Quality Control Charts. *Quality and Reliability Engineering International*, 25:773–804, 2009.

[193] Abraham Wald. *Sequential Analysis*. John Wiley and Sons, 1st edition, 1947.

[194] Dashun Wang, Z. Wen, Hanghang Tong, C.Y. Lin, C. Song, and A.L. Barabási. Information spreading in context. In *Proceedings of the 20th international conference on World wide web*, pages 735–744. ACM, 2011.

[195] Yao Wang and Julita Vassileva. Toward Trust and Reputation Based Web Service Selection: A Survey. *International Transactions on Systems Science and Applications*, 3(2):118–132, 2007.

[196] Yonghong Wang and Munindar P Singh. Formal Trust Model for Multiagent Systems. In *Proceedings of the 20th International Joint Conference on Artifical intelligence*, pages 1551–1556, 2007.

[197] Yonghong Wang and Munindar P. Singh. Evidence-based trust: A mathematical model geared for multiagent systems. *ACM Transactions on Autonomous and Adaptive Systems (TAAS)*, 5(4):14, 2010. ISSN 1556-4665.

[198] Yonghong Wang, Chung-Wei Hang, and Munindar P. Singh. A Probabilistic Approach for Maintaining Trust Based on Evidence. *Journal of Artificial Intelligence*, 40:221–267, 2011.

[199] Larry Wasserman. An Inferential Interpretation of Default Priors. Technical report, Carnegie-Mellon University, 1991.

[200] Eugene Webb. Trust and Crisis. In R. M. Kramer and T. R. Tyler, editors, *Trust in Organizations: Frontiers of Theory and Research*, pages 288–301. Sage, Thousand Oaks, 1996.

[201] K. E. Weick and K. H. Roberts. Collective Mind in Organizations: Heedful Interrelating on Flight Decks. *Administrative Science Quarterly*, 38:357–381, 1993.

[202] Andrew Whitby, Audun Jøsang, and Jadwiga Indulska. Filtering out Unfair Ratings in Bayesian Reputation Systems. *The ICFAIN Journal of Management Research*, 4(2):48–64, 2005.

[203] Frank Wilcoxon. Individual comparisons by ranking methods. *Biometrics*, 1:80–83, 1945.

[204] Edwin Bidwell Wilson. Probable Inference, the Law of Succession, and Statistical Inference. *Journal of the American Statistical Association*, 22:209–212, 1927.

[205] T. Yamagishi and M. Yamagishi. Turst and Commitment in the United States and Japan. *Motivation and Emotion*, 18(2):129–166, 1994.

[206] Yafei Yang, Yan (Lindsay) Sun, and Steven Kay. Defending online reputation systems against collaborative unfair raters through signal modeling and trust. In *Proceedings of the 2009 ACM Symposium on Applied Computing (SAC)*, 2009.

[207] H. Yu, Zhiqi Shen, Chunyan Miao, Cyril Leung, and Dusit Niyato. A Survey of Trust and Reputation Management Systems in Wireless Communications. *Proceedings of the IEEE*, 98(10):1755–1772, 2010. ISSN 0018-9219.

[208] Mauricio Zambrano-Bigiarini. *R-Package hydroGOF: Goodness-of-fit functions for comparison of simulated and observed hydrological time series*, 10 2012. URL http://cran.r-project.org/web/packages/hydroGOF/hydroGOF.pdf.

COLOPHON

This document was typeset using the typographical look-and-feel `classicthesis` developed by André Miede. The style was inspired by Robert Bringhurst's seminal book on typography "*The Elements of Typographic Style*". `classicthesis` is available for both LaTeX and LyX:

http://code.google.com/p/classicthesis/

Happy users of `classicthesis` usually send a real postcard to the author, a collection of postcards received so far is featured here:

http://postcards.miede.de/

Final Version as of October 1, 2015 (`classicthesis` version 4.0).

ERKLÄRUNG

Hiermit erkläre ich, die vorgelegte Arbeit zur Erlangung des akademischen Grades *Dr. rer. nat.* mit dem Titel

On the Statistics of Trustworthiness Prediction

selbständig und ausschließlich unter Verwendung der angegebenen Hilfsmittel erstellt zu haben. Ich habe bisher noch keinen Promotionsversuch unternommen.

Darmstadt, October 1, 2015

Sascha Hauke

WISSENSCHAFTLICHER WERDEGANG

10/2002 – 07/2007	Studium der Informatik, Anwendungsfach Sprachwissenschaften, Westfälische Wilhelms-Universität Münster Abschluss: Diplom-Informatiker Diplomarbeitsthema: *Modelling and Validating the Trust Establishment Protocol FEDUST*, in Kooperation mit: BMW Forschung and Technik GmbH, München (eine Tochterfirma der BMW Group)
01/2008 – 04/2009	Mitarbeiter, IZKF Forschungsgruppe 4, Universitätsklinikum Münster
05/2009 – 10/2010	Wissenschaftlicher Mitarbeiter, Arbeitsgruppe Optimierte Systeme, Fachhochschule der Wirtschaft (FHDW) in Bergisch Gladbach
seit 11/2010	Wissenschaftlicher Mitarbeiter, Fachgebiet Telekooperation, Technische Universität Darmstadt